FACULTAD DE ESTUDIOS ESTADÍSTICOS

Master de Minería de Datos e Inteligencia de Negocios.

Curso 2015/2016

Trabajo de Fin de Master

TITULO: Metodología de Minería de datos contra el fraude empresarial.

***Alumno*: Daniel Martín García.**

***Tutor*: Javier Castro Cantalejo.**

Junio de 2016

UNIVERSIDAD COMPLUTENSE
MADRID

Agradecimientos:

A mis padres y hermano por aguantarme, a Javier Castro y Rosa por ayudarme y guiarme siempre y a Javier Portela, gran docente, siempre dispuesto a ayudar al alumnado.

Índice

1. Introducción.

Por motivos de confidencialidad con la empresa suministradora de los datos, los cuales serán el objeto de análisis en el siguiente estudio, se avisa que para garantizar dicha confidencialidad, los datos han sido "camuflados", situándolos en un contexto diferente aunque de similar comportamiento al original, por ello parte de las variables, así como las categorías de muchas de ellas aparecerán con nombres codificados. Esto supondrá a veces problemas para generar conclusiones o interpretaciones.

No hacen falta sofisticados estudios o encuestas para conocer la magnitud del fraude fiscal existente en España porque a diario se puede comprobar la alegría con la que se ofrecen en el mercado servicios u operaciones sin factura, facturas sin IVA, u operaciones con una parte en negro. Esta situación pone de manifiesto el poco miedo que tiene el defraudador a ser pillado, la falta de repudio social ante el fraude fiscal.

Aunque existe un gran campo de mejora se lleva trabajando tiempo atrás contra la lucha del fraude, ya que la propia Unión Europea está tomando conciencia de la gravedad de este asunto. Así, el Parlamento Europeo, en un informe de 17 de julio de 2012, se lamenta de que no se disponga de cifras precisas. Las estimaciones sobre las pérdidas globales (directas e indirectas) de ingresos fiscales originadas por el fraude fiscal se sitúan en Europa entre los 200.000 y los 250.000 millones de euros anuales.

Existen informes, estudios o datos que ponen de manifiesto la magnitud del problema y que, en términos porcentuales, cuantifican el fraude en España entre el 20% y 25%, el doble que la media de la UE. Un ejemplo es el informe *El huevo que deja el diablo: Una estimación del fraude en el IRPF con microdatos tributarios* publicado en 2014 por la empresa *Fedea.*

Podemos señalar que el fraude fiscal no es un fenómeno nuevo, sino que ha supuesto un problema importante en la sociedad española en las últimas décadas. El PIB de España alcanzó en el primer trimestre de 2015 los 270.703 millones de euros, y la presión fiscal media se sitúa en el 40.71%; la cifra de cuotas no ingresadas supera los 70.000 millones de euros cada año como mínimo. De acuerdo con los datos anteriores, y teniendo en cuenta que el periodo de prescripción establecido en la Ley General Tributaria es de cuatro años, se puede cuantificar en una cifra aproximada de 280.000 millones de euros las cuotas tributarias no ingresadas, las cuales, es posible su ingreso.

Desde su creación, la Agencia Tributaria ha elaborado diversos documentos de planificación estratégica en 1994, 1996, 1998 y 2000, que pueden citarse como antecedentes del Plan de Prevención del Fraude aprobado en marzo de 2010 y que todavía está ejecutándose.

2. Naturaleza de los datos.

Los datos objeto del análisis en el presente estudio recogen información temporal sobre la totalidad de empresas registradas en la Agencia Tributaria Española. La temporalidad de los datos es trimestral, siendo los primeros datos del primer trimestre del año 1997 y los últimos del segundo trimestre del año 2014, es decir, se tiene información de un total de 66 trimestres.

La información recogida por la Agencia Tributaria es muy variada y completa; de cada empresa registrada se recopila tanto información subyacente a dicha empresa (situación geográfica, sector económico, beneficios declarados, reclamaciones presentadas, deudas, etc.) como información generada por la propia Agencia Tributaria (Anomalías observadas, inspecciones, beneficios estimados, etc.).

La fuente de información del presente estudio está integrada por más de 19 millones y medio de registros resultado de sumar los registros de cada uno de los 66 trimestres de los que se tiene información; el número de registros por trimestre presenta una tendencia creciente que pasa de aproximadamente 260 mil registros a comienzos del año 1997 a 320.000 registros aproximadamente a mediados de 2014.

Cabe mencionar que la información de partida precisó de un amplio proceso de depuración, puesto que dicha información venía desde diferentes departamentos de la Agencia Tributaria, donde en cada uno de ellos se publicaban los datos en formatos muy distintos. La información recibida por cada departamento era información mensual donde ni siquiera se respetaban los formatos entre los diferentes meses. Todo esto obligo a realizar como ya se ha comentado un amplio proceso de depuración donde se tuvo que estar en constante comunicación con la Agencia Tributaria.

La depuración realizada consistió principalmente en unificar los formatos de las múltiples tablas recibidas por cada departamento de la Agencia Tributaria, no obstante también se precisó de procesos como estimación de valores perdidos o missing, acotaciones de variables o la realización de infinidad de cruces de tablas.

3. Objetivos y Metodología

Debido a la gran cantidad de información contenida en los datos proporcionados por la Agencia Tributaria, es necesario fijar a priori los objetivos fundamentales del mismo y la metodología necesaria para su obtención.

El objetivo principal del estudio es la predicción clasificatoria de las empresas registradas ante la Agencia Tributaria. Se entenderá como fraude, el hecho de que una empresa declare ante la Agencia Tributaria un beneficio obtenido inferior al real.

Por otro lado surgen objetivos específicos u objetivos intermedios, los cuales, principalmente se establecen para la búsqueda del objetivo principal del estudio, alguno de los más destacados son:

- Conocer a fondo la información de partida a partir de un profundo análisis Descriptivo.
- Reducir las dimensiones de las variables independientes del estudio.
- Establecer una metodología clara de las técnicas de predicción utilizadas.
- Eliminar la dimensión temporal de nuestras variables con la creación de nuevas variables tratadas.
- Dividir la base de datos en tres partes (Aprendizaje, Validación y Test) con la finalidad de conseguir modelos más óptimos en cuanto a su validez predictiva.
- Encontrar los mejores parámetros estructurales para las diferentes técnicas de predicción utilizadas.
- Estableces una comparativa que permita diferenciar el mejor modelo predictivo creado.
- Realizar un pos-análisis que nos ayude a entender en la medida de lo posible el comportamiento del que sea seleccionado como mejor modelo de predicción.

Las técnicas estadísticas multivariantes que se van a utilizar para llevar a cabo este objetivo son las siguientes:

Análisis Factorial; se utilizará para disminuir el número de variables explicativas a utilizar en la predicción probabilística.

Regresión Logística, Redes Neuronales, Random Forest, Gradient Boosting y Ensamble de Modelos; se utilizarán para la predicción probabilística de que una empresa este cometiendo fraude.

A continuación se presenta un breve resumen de la metodología SEMMA que se intenta seguir en el presente estudio así como un breve resumen de cada una de las técnicas estadísticas anteriormente citadas.

3.1. Metodología SEMMA.

Para alcanzar los objetivos establecidos se ha procurado seguir el patrón de trabajo impuesto por la Metodología SEMMA, la cual responde al siguiente diagrama.

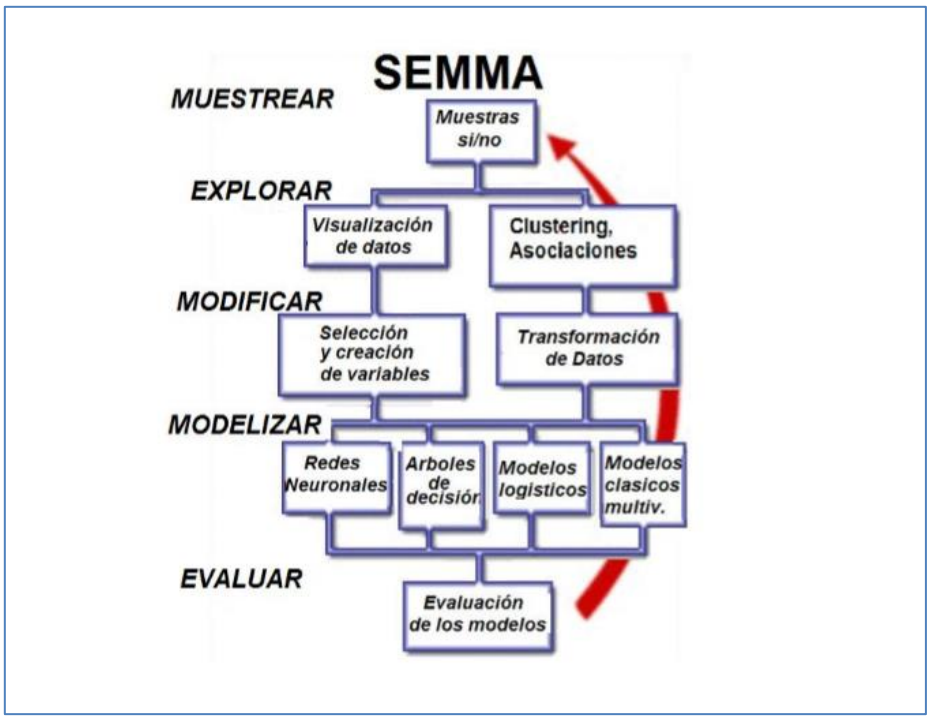


Diagrama 3.1. Metodología SEMMA

El esquema anterior no ha de tomarse al pie de la letra: ni el orden, ni el contenido.

- No siempre intervienen todas las fases del proceso (por ejemplo, muestrear en el presente estudio no se lleva acabo).
- A menudo el orden no es exacto (se puede explorar antes de muestrear; se puede explorar-modificar-explorar-modificar, etc.).
- El proceso se suele repetir muchas veces, pasando de unas fases a otras sin respetar el orden del flujo.

3.2. Análisis factorial.

El Análisis Factorial es una técnica estadística de reducción de datos usada para explicar las correlaciones entre las variables observadas en términos de un número menor de variables no observadas llamadas factores. Para ello utiliza un conjunto de variables aleatorias inobservables, que llamaremos factores comunes, de forma que todas las covarianzas o correlaciones son explicadas por dichos factores y cualquier porción de la varianza inexplicada por los factores comunes se asigna a términos de error residuales que llamaremos factores únicos o específicos.

El análisis factorial es, por tanto, una técnica de reducción de dimensionalidad de los datos. Su propósito último consiste en buscar el número mínimo de dimensiones capaces de explicar el máximo de la información contenida en los datos.

A diferencia de lo que ocurre en otras técnicas como el análisis de la varianza o el de regresión, en el análisis factorial todas las variables cumplen el mismo papel: todas ellas son independientes en el sentido de que no se pretende explicar unas en función de otras.

El análisis factorial consta de cuatro fases características: el cálculo de una matriz capaz de expresar la variabilidad conjunta de todas las variables, la extracción del número óptimo de factores, la rotación de la solución para facilitar su interpretación y la estimación de las puntuaciones de los sujetos en las nuevas dimensiones.

Formalmente, sean X_1, X_2, …, X_p las p variables objeto de análisis, que supondremos en todo lo que sigue que están tipificadas, de no estarlo el análisis se realizaría de forma similar pero la matriz utilizada para el cálculo de los factores no sería la matriz de correlaciones sino la de varianzas-covarianzas. Se parte de la matriz de datos

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Con el objetivo de llegar al modelo del análisis factorial que viene dado por las ecuaciones:

$$X_1 = a_{11}f_1 + a_{12}f_2 + \ldots + a_{1k}f_k + u_1$$
$$X_2 = a_{21}f_1 + a_{22}f_2 + \ldots + a_{2k}f_k + u_2$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$X_p = a_{p1}f_1 + a_{p2}f_2 + \ldots + a_{pk}f_k + u_p$$

Siendo la matriz |A| = $\begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pk} \end{pmatrix}$ la matriz de puntuaciones de cada factor con

respecto a las variables originales, $f_1,\ldots,f_k$, los factores o variables no observables que se buscan y $u_1,\ldots,u_k$, términos de error independientes e idénticamente distribuidos.

Se supone, además, que los factores comunes están a su vez estandarizados ($E(F_i)=0$; $Var(F_i)=1$), los factores específicos tiene media 0 y están incorrelados ($E(u_i)=0$; $Cov(ui,uj)= 0$ si $i\neq j$; j, i=1,…,p) y que ambos tipos de factores también están incorrelados entre sí ($Cov(F_i,u_j)=0$, para todo i=1,….,k; j=1,…,p) .

Si además, los factores comunes también son incorrelados ($Cov(F_i,F_i)=0$, si $i\neq j$; j, i=1,…,k) estamos ante un modelo con factores ortogonales (caso más ideal).

Los contrastes utilizados en el análisis factorial son los siguientes:

La medida de adecuación muestral *KMO (Kaiser-Meyer-Olkin)* contrasta si las correlaciones parciales [(1)] entre las variables son suficientemente pequeñas. Permite comparar la magnitud de los coeficientes de correlación observados con la magnitud de los coeficientes de correlación parcial. El estadístico KMO varía entre 0 y 1. Los valores pequeños indican que el análisis factorial puede no ser una buena idea, dado que las correlaciones entre los pares de variables no pueden ser explicadas por otras variables. Los menores que 0.5 indican que no deben utilizarse el análisis factorial con los datos que se han utilizado.

La *prueba de esfericidad de Bartlett* contrasta la hipótesis nula de que la matriz de correlaciones es una matriz identidad, en cuyo caso no existirían correlaciones significativas entre las variables y el modelo factorial no sería pertinente.

3.2. Regresión Logística.

Los modelos de regresión logística son modelos estadísticos en los que se desea conocer la relación entre: Una variable dependiente cualitativa (en nuestro caso dicotómica: No cometer fraude (codificado como menos 1) y cometer fraude (codificado como 1)) y una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas, siendo la ecuación inicial del modelo de tipo exponencial. Las covariables cualitativas deben ser dicotómicas y si una covariable cualitativa tuviera más de dos categorías, para su inclusión en el modelo se deberá realizar una transformación de la misma en varias covariables cualitativas dicotómicas ficticias o de diseño (las llamadas

variables dummy), de forma que una de las categorías se tomaría como categoría de referencia. Con ello, cada categoría entraría en el modelo de forma individual. En general, si la covariable cualitativa posee n categorías, habrá que realizar n−1 covariables ficticias. Por otro lado las variables cuantitativas no precisan de un tratamiento previo para su introducción en los modelos de regresión Logísticos. Por sus características, los modelos de regresión logística permiten dos finalidades:

Por un lado, cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente, lo que lleva implícito también clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente (es decir, conocer la odds ratio para cada covariable). Entendiendo por odds ratio una medida que cuantifica el cambio de probabilidad que se produce por el cambio en la categoría de cada variable respecto a la categoría marcada como referencia en el caso de variables categóricas o dicotómicas. En el caso de variables continuas, el OR es la medida que presenta el cambio en la probabilidad logística al aumentar una unidad la variable predictiva, es decir, nos da una tendencia general del comportamiento de la variable respuesta frente a la variable del modelo.

Por otro lado, clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables.

En resumen el objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos, es decir, el objetivo consiste en determinar la $P[Y=1/X_1,X_2,\ldots,X_K]$.

Para ello, se construye el modelo $P[Y=1/X_1,X_2,\ldots,X_K] = G(X_1,X_2,\ldots,X_K;\beta)$ donde:

$G(X_1,X_2,\ldots,X_K;\beta)$ es la función que va de los reales al intervalo [0,1] que depende de un vector de parámetros $\beta=(\beta_1, \beta_2,\ldots, \beta_k)$.

El método de estimación clásico utilizado por la regresión logística y en el presente estudio es el de Máxima Verosimilitud, el cual, proporciona estimadores consistentes, asintóticamente normales y eficientes.

$$G(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}$$

Las principales ventajas de la regresión logística son:

- El análisis de regresión logística es una herramienta muy flexible en cuanto a la naturaleza de las variables explicativas, pues éstas pueden ser numéricas y categóricas.

- El modelo de regresión logística es robusto con respecto al incumplimiento moderado del supuesto de igualdad de las matrices de covarianza entre grupos (heterocedasticidad).

- La regresión logística no hace supuestos sobre la distribución de las variables independientes, ya que en el modelo estas no son consideradas como variables aleatorias.

- Tamaño de muestra y número de variables independientes. Una de las ventajas de la regresión logística es que permite el uso de múltiples variables con relativamente pocos casos.

La principal desventaja de la regresión logística es:

- La multicolinealidad entre las variables, traerá como consecuencia grandes errores estándar y coeficientes estimados anormalmente elevados.

3.3. Redes Neuronales.

Las Redes Neuronales constituyen una herramienta muy potente de análisis, modelización y predicción. Se rigen por la filosofía general de obtener modelos coherentes con la realidad observada, de tal modo que sean los datos los que determinen el comportamiento de la red, ya sea a través de la determinación de sus estructuras o de sus parámetros internos. Esta técnica forma parte de los métodos no paramétricos de análisis de datos.

Una red neuronal consiste en un conjunto de unidades de procesamiento, conocidas como nodos o unidades, las cuales están conectadas entre sí. La conectividad de una red neuronal viene dada en términos de una arquitectura, la cual es un grafo con conexiones entre los nodos. Aquellos nodos que no tienen conexiones de entrada se denominan nodos de entrada y los nodos de los que no sale ninguna conexión se denominan nodos de salida. El resto de nodos se denominan nodos ocultos. Los nodos de computación de una red son los nodos de salida y los nodos ocultos. Todos los nodos que se encuentran a la misma distancia en el grafo de los nodos de entrada forman una capa. Cada nodo i-ésimo de una red neuronal está caracterizado por un valor numérico denominado valor o estado de activación y_i asociado a la unidad. Existe una función de salida o de activación f_i que transforma el estado actual de activación en una señal de salida $f_i(y_i)$.

La composición de la activación de la unidad y la función de salida se denomina función de transferencia de la unidad. Las funciones de activación son en general diferentes según se trate de unidades de las capa de salida o de unidades pertenecientes a la capa oculta.

Una red neuronal (NN) es un sistema de cálculo hecho con una cantidad de elementos de procesado interconectados. Las dos componentes primarias de una red neuronal son los elementos de procesado (los nodos equivalentes a las neuronas biológicas) y sus interconexiones.

La estructura de una red neuronal *Perceptrón Multicapa* como ya hemos comentado, se dispone en capas: de datos de entrada, de nodos ocultos y de datos de salida:

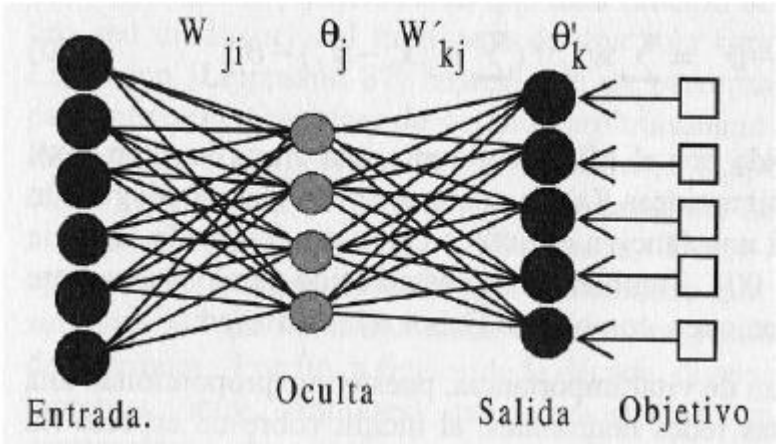


Gráfico 3.3.1.

La capa de entrada se conecta con la capa oculta (θ_j) mediante una función de combinación, donde los pesos w_{ij} *(pesos* sinápticos*)* hacen el papel de parámetros a estimar. Sobre esta función se aplica una función de activación[(2)], que pueden ser entre otras: función sigmoidal, función gaussiana, función tangente hiperbólica, etc. De los nodos ocultos a los nodos de salida (θ'_k) se aplica el mismo procedimiento sobre las nuevas variables provenientes de los nodos ocultos: una función de combinación y ocasionalmente una de activación. El valor final de la función de activación en cada nodo oculto es el valor de salida en ese nodo.

Las Redes Neuronales tienen numerosas ventajas frente a otras técnicas de predicción, describiendo las más destacables a continuación:

El aprendizaje del modelo no necesita ser programado, las redes neuronales son capaces de extraer sus propias reglas a partir de ejemplos reales mediante la adaptación de la matriz de ponderaciones. Estas reglas quedan almacenadas y extendidas a lo largo de las conexiones.

Otra ventaja es que son tolerantes al ruido, es decir, son capaces de abstraer las características esenciales de los datos y así generalizar de forma correcta aún en presencia de datos distorsionados o incompletos.

No son paramétricas, no necesitan hacer supuestos de la forma funcional de la función que van a aproximar, ni sobre la distribución de las variables independientes.

Además no tienen por qué ser lineales, permiten realizar a través de sus funciones de activación todo tipo de transformaciones de los datos, lo cual supone una gran ventaja frente a los modelos tradicionales de regresión lineal, logística o discriminante.

Pero tiene dos grandísimas limitaciones, por un lado la imposibilidad de determinar cómo se procesa internamente la información y por otro, que no existe aún una metodología clara y rigurosa para determinar el número de capas ocultas o el número de nodos que tiene que tener cada capa, lo que hace difícil encontrar el modelo óptimo a la primera, se trata más bien de un proceso de ensayo-error del investigador.

Los software estadísticos solventan o intentan solventar este último inconveniente sugiriendo posibles modelos de Redes Neuronales o probando múltiples de ellos, proporcionado el óptimo.

3.4. Arboles de Clasificación.

La técnica de predicción "Arboles de Clasificación" no se aplicara de forma directa para el alcance de los objetivos del presente estudio, no obstante, otros algoritmos (Random Forest y Gradient Boosting), se basan en diferentes modificaciones de dicha técnica. Por ello se considera necesaria su explicación teórica.

Un árbol de clasificación es el resultado de preguntar una secuencia ordenada de cuestiones, estando el tipo de cuestiones que se preguntan en cada etapa en dependencia con las respuestas a las cuestiones previas de la secuencia.

El punto único de inicio del árbol de clasificación se llama *nodo raíz* y contiene el conjunto total a clasificar 'ζ' en la parte superior del árbol. Un *nodo* es un subconjunto del conjunto de variables, y puede ser terminal o no terminal. Un *nodo no terminal* o *padre* es un nodo que se divide en *'k'* nodos descendientes, en el presente estudio se utilizarán los arboles de decisión binarios en los que cada *nodo padre* se divide en dos *nodos descendientes*.

Una división binaria queda determinada mediante una condición booleana sobre los valores de una única variable, siendo satisfecha la condición ('si') o no satisfecha ('no')

por el valor observado de dicha variable. Todas las observaciones de 'ζ' que alcanzan un nodo (padre) particular y satisfacen la condición para dicha variable van a parar a uno de los nodos descendientes; el resto de las observaciones de dicho nodo (padre) que no satisfacen la condición van a parar al otro nodo descendiente.

Un nodo que no se divide se llama *nodo terminal* y se le asigna una etiqueta de clase. Cada observación de 'ζ' va a parar a uno de los nodos terminales. Cuando una observación de clase desconocida transita a través del árbol y va a parar a un nodo terminal, se le asigna la clase correspondiente a la etiqueta de clase adjunta a dicho nodo. Puede haber más de un nodo terminal con la misma etiqueta de clase.

En cada nodo, el algoritmo de generación del árbol tiene que decidir sobre qué variable es 'óptima' la partición en dos. Necesitamos considerar cada posible división sobre todas las variables presentes en dicho nodo, enumerar después todas las posibles divisiones, evaluar cada una, y decidir cuál es la mejor siguiendo algún criterio.

En el presente estudio se construirá el árbol de tal forma que se procure construir nodos homogéneos en cuanto a la variable dependiente.

Antes de elegir la mejor división sobre todas las variables necesitamos saber elegir la división óptima para una variable dada. Para ello definimos una medida de la bondad de una división.

Sean $\prod_1,\ldots,\prod_k$, k > 2 las clases a las que da lugar una variable dada sobre el nodo *t*, definimos la *función de impuridad* del nodo *i(t)* como sigue:

$$i(t) = \varphi(p(1|t),\ldots,p(k|t))$$

donde p(k|t) es un estimador de p(X pertenezca a $\prod_k|t$). La probabilidad condicional de que una observación **X** esté en $\prod_k$ dado que se observa en el nodo t. En la expresión de *i(t)* imponemos que φ sea una función simétrica definida para el conjunto de todas las k-uplas de probabilidades $(p_1,\ldots,p_k)$ cuya suma es la unidad, que se minimice en los puntos (1, 0, …, 0), (0, 1, …, 0), …, (0, 0, …, 1), y que se maximice en el punto $(\frac{1}{k},\ldots,\frac{1}{k})$.

Una de las funciones más utilizadas para φ, la cual, es también la que se utilizará en el presente estudio es el índice de diversidad de Gini cuya expresión es:

$$i(t) = 1-p^2-(1-p)^2 = 2p(1-p)$$

Donde *p* es la probabilidad de que una observación este bien clasificada en el nodo *t*.

Adoptaremos la estrategia de dejar que el árbol crezca hasta que cada nodo contenga menos de un valor mínimo de observaciones al que notaremos como n_{min} (valor escogido por el investigador) y después podaremos las ramas del árbol hasta que el árbol tenga un 'tamaño adecuado'. Un árbol podado es un subárbol del árbol original mayor.
A continuación se muestra un ejemplo sencillo sobre arboles de clasificación, en el cual, el objetivo es predecir una variable con tres categorías a partir de otro conjunto de variables numéricas:

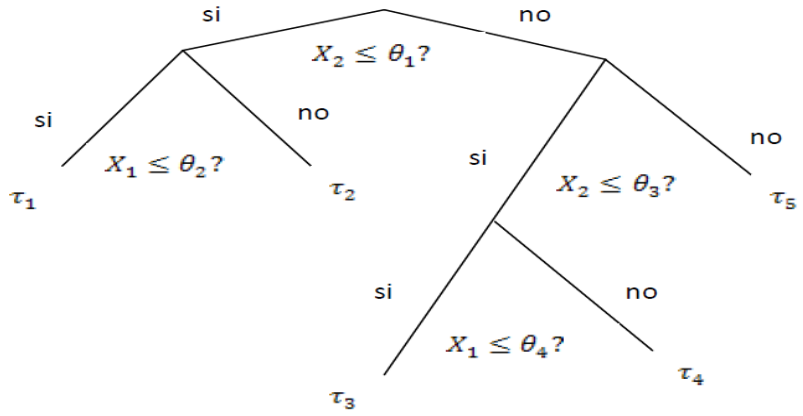


Gráfico 3.4.1.

El árbol presentado tiene una profundidad igual a 3, entendiendo por profundidad al máximo de preguntas que se deben contestar para llegar a un nodo terminal u hoja. Está formado por 9 nodos, 5 de ellos son nodos terminales $(t_i)_{i=1,2,3,4,5}$, 3 nodos padres y el restante (el primero) nodo raíz, a partir del cual, respondiendo a una serie de preguntas o condiciones booleanas se llega a un nodo hoja al cual se le asigna una categoría, siendo los valores de θ_i números pertenecientes a los reales que hacen de frontera para la clasificación en un nodo u otro en cada decisión booleana, por último x_i representan a las variables explicativas.

La principal ventaja que ofrecen los arboles de clasificación frente a otras técnicas estadísticas de predicción es que sus reglas de asignación son legibles, lo cual, produce una sencilla e intuitiva interpretación. Otras características que hacen llamativa a esta técnica son que es una técnica no paramétrica que tiene en cuenta las interacciones que pueden existir entre los datos, es robusta frente a datos atípicos o individuos mal etiquetados, es válida sea cual sea la naturaleza de las variables explicativas: continuas, nominales u ordinales.

Pero no todo son ventajas, esta técnica también presenta importante inconvenientes. Una gran desventaja para el investigador es la dificultad para elegir el árbol óptimo. Por otro lado, las reglas de asignación son bastantes sensibles a pequeñas perturbaciones en

los datos (inestabilidad). Los árboles de clasificación requieren un gran número de datos para asegurarse que la cantidad de las observaciones de los nodos hoja es significativa.

3.5. Random Forest.

Random Forests (Breiman, 2001)[3] es un algoritmo para clasificación y regresión de amplio uso en la comunidad que tiene un rendimiento bueno para datos de alta dimensionalidad. Random Forest puede considerarse como un derivado de los Árboles de Clasificación.

El esquema del algoritmo Random Forest es:

Dados los datos de tamaño N.

1. Repetir m veces i), ii), iii):

   i) Seleccionar n<N observaciones con reemplazamiento de los datos originales.

   ii) Aplicar un árbol de la siguiente manera:
   - En cada nodo, seleccionar p variables de las k originales y de las p elegidas, escoger la mejor variable para la partición del nodo.

   iii) Obtener predicciones para todas las observaciones originales N.

2. Promediar las m predicciones obtenidas en el apartado 1).

Los principales parámetros a controlar en Random Forest son:

- El tamaño de las muestras ‘n’ y si se va a utilizar bootstrap (con reemplazo) o sin reemplazamiento.
- El número de iteraciones ‘m’ a promediar
- El número de variables p a muestrear en cada nodo (si es igual al número inicial de variables ‘k’ el Random Forest es equivalente a otra técnica de predicción denominada “Bagging”)

- Características de los árboles. Son bastante influyentes:

    - El número de hojas final o, en su defecto, la profundidad del árbol.
    - El maxbranch (número de divisiones máxima en cada nodo. Por defecto se dejará en 2, árboles binarios).
    - El p-valor para las divisiones en cada nodo. Más alto -> árboles menos complejos (más sesgo, menos varianza).
    - El número de observaciones mínimo en una rama-nodo. Se puede ampliar para evitar sobreajuste (reducir varianza) o reducir para ajustar mejor (reducir sesgo).

Random Forest trata de incorporar dos fuentes de variabilidad a los Árboles de Clasificación (remuestreo de observaciones y de variables) para ganar en capacidad de generalización, y reducir el sobreajuste conservando a la vez la facultad de ajustar bien relaciones particulares en los datos (interacciones, no linealidad, cortes, problemas de extrapolación, etc.).

Las principales ventajas de esta técnica frente a los Árboles de Clasificación son:

- Aumenta la capacidad predictiva y disminuye la varianza.
- Disminuye la sensibilidad frente a cambios en los datos, aumenta la estabilidad y la robustez.
- Aumenta la suavidad (función menos escalonada), lo que a veces redunda en menor error promedio de predicción.

Por otro lado, esta técnica de predicción presenta la desventaja de la perdida de interpretabilidad de los resultados, donde solo se puede evaluar la importancia de cada una de las variables explicativas del modelo. Creándose un ranking de las variables según su frecuencia utilizada en el algoritmo.

2.6. Gradient Boosting.

El algoritmo gradient boosting consiste en repetir la construcción de árboles de regresión/clasificación, modificando ligeramente las predicciones iniciales cada vez, intentando ir minimizando los residuos en la dirección de decrecimiento.

Al plantear diferentes árboles cada vez, el proceso va ajustando las predicciones cada vez más a los datos, y de alguna manera unos árboles corrigen a otros con lo cual la flexibilidad y adaptación del método mejora respecto a la construcción de un único árbol.

Este proceso ha de ser monitorizado en principio mediante early stopping[6] para determinar el número de iteraciones. Por lo tanto necesitará datos de validación.

Aunque a menudo el early stopping no es necesario pues la convergencia es lenta y van a la par los errores en training y validación (no sobreajuste).

El algoritmo Gradient Boosting utiliza la función logit como función base, y la Deviance[7] como función de error. El objetivo es ir retocando la función logit para que en cada paso del algoritmo se actualicen las probabilidades predichas y los residuos:

Se define L(yi,f(xi)) = log(1+$e^{-2yif(xi)}$), donde yi = 1, 0.

Y f(xi) se define como f(xi)=0.5 log ($\frac{\hat{p}_i^{(m}}{1-\hat{p}_i^{(m}}$), con pi=p(yi=1).

Pasos a seguir en el algoritmo:

1) $\hat{p}_i^{(0}$=% de 1 en los datos.

2) Calcular el residuo actual $\hat{r}_i^{(m}$=yi-$\hat{p}_i^{(m}$ (este residuo es el gradiente, dada la función de error Deviance).

3) Ajustar mediante un árbol de regresión los residuos $\hat{r}_i^{(m}$= variable dependiente, X -> vector de variables predictoras -> $\hat{r}_i^{(m)}$.

4) Actualizar fi mediante $\hat{f}_i^{(m+1}$ = $\hat{f}_i^{(m}$ + v. $\hat{r}_i^{(m}$= 0.5 log(1+$e^{-2yif(xi)}$). v. $\hat{r}_i^{(m}$. Donde "v" es un parámetro de regularización encargado de "suavizar" o definir la importancia de cada predicción.

5) Actualizar la probabilidad predicha mediante $\hat{p}_i^{(m+1}=\frac{1}{1+e_i^{-2f_i^{(m+1)}}}$

6) Volver al paso 2).

Las principales ventajas de esta técnica son:

- Invariante frente a transformaciones monótonas: no es necesario realizar transformaciones logarítmicas, etc.
- Buen tratamiento de missing, variables categóricas, etc. Universalidad.
- Muy fácil de implementar, relativamente pocos parámetros a monitorizar (número de hojas o profundidad del árbol, tamaño final de hojas, parámetro de regularización…).
- Gran eficacia predictiva, algoritmo muy competitivo. Supera a menudo al algoritmo Random Forest.
- Robusto respecto a variables irrelevantes. Robusto respecto a multicolinealidad.
- Detecta interacciones ocultas.

Al igual que pasaba con el algoritmo de Random Forest y con todos los algoritmos derivados de la creación de Árboles de Clasificación en masa se pierde la interpretabilidad de los resultados, donde solo se puede evaluar la importancia de cada una de las variables explicativas del modelo. Creándose un ranking de las variables según su frecuencia utilizada en el algoritmo.

2.7. Ensamble de Modelos.

Los métodos Ensamble consisten en la construcción de predicciones a partir de la combinación de varios modelos. Existen infinidad de métodos para la combinación de las distintas predicciones.

Se proponen aquí algunos métodos de ensamble de los modelos que se van a llevar a cabo en el presente estudio con el fin de obtener clasificadores cuya optimización de clasificación sea mayor que la proporcionada por los modelos individuales. Para ello se construye un conjunto de datos que contiene las probabilidades estimadas por los diferentes modelos ajustados con el fin de combinar estas probabilidades de distintas formas. Sin ánimo de profundizar mucho en los modelos de ensamble óptimos, se

proponen varios de estos posibles clasificadores combinados y se comparara su capacidad.

En primer lugar se construyen diferentes clasificadores a partir de diferentes combinaciones de las probabilidades estimadas. Se llevara a cabo todas las combinaciones posibles de modelos. La forma de agregar la información de las diferentes predicciones se llevará a cabo mediante diferentes criterios: El mínimo de las predicciones (todos los modelos deben considerar que existe fraude), el máximo de las predicciones (al menos uno de los modelos considera que existe fraude), la mediana (la mayoría de los modelos lo consideran la existencia de fraude) y por último la media (en media los modelos consideran la existencia de fraude).

A continuación se realizará un ajuste de regresión logística por pasos, Stepwise, con las diferentes probabilidades estimadas como predictores para la clasificación del evento y se construirá un clasificador que viene dado por la media ponderada por pesos obtenidos por los coeficientes de la regresión obtenidos, de forma relativa. Así mismo se considerará la probabilidad estimada de este modelo de regresión logística como otro posible ensamble logístico.

Las principales ventajas del ensamble de modelos son:

- Bastante robustos, unos modelos corrigen a otros.
- Reducen la varianza del error en general, casi nunca empeoran los modelos.

Por otro lado se comentan las principales desventajas de los métodos ensamble:

- Cada modelo tiene sus errores de estimadores de parámetros lo que aumenta aparentemente la complejidad.
- Excesivas posibilidades que a veces llevan al sobreajuste.
- Los resultados no son interpretables.

2.8. Comparación de modelos.

Para la comparación entre modelos se utilizará el coeficiente de correlación Matthews (MCC). Dicho coeficiente de correlación se utiliza en la máquina de aprendizaje como una medida de la calidad de las clasificaciones binarias, introducido por el bioquímico Brian W. Matthews en 1975[(4)].

El MCC es en esencia un coeficiente de correlación entre las clasificaciones binarias observadas y predichas; devuelve un valor entre -1 y +1. Un coeficiente de 1 representa

una predicción perfecta, 0 no es mejor que la predicción aleatoria y -1 indica total desacuerdo entre la predicción y la observación.

MCC está relacionado con el estadístico Chi-cuadrado para una tabla de contingencia de 2x2

$$|\text{MCC}| = \sqrt{\frac{X^2}{n}},$$

donde n es el número total de observaciones. Aunque no hay una manera perfecta de describir la matriz de confusión de los aspectos positivos y negativos verdaderos y falsos por un solo número, el coeficiente de correlación Matthews es generalmente considerado como uno de los mejores de tales medidas[5].

El MCC se puede calcular directamente de la matriz de confusión utilizando la fórmula:

$$\text{MCC} = \frac{VP\ x\ VN - FP\ x\ FN}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}}$$

En esta ecuación, VP es el número de verdaderos positivos , VN el número de verdaderos negativos , FP el número de falsos positivos y FN el número de falsos negativos . Si cualquiera de los cuatro sumas en el denominador es cero, el denominador se puede establecer arbitrariamente a uno.

No obstante para considerar que modelo es mejor que otro también se tendrá en cuenta el número de parámetros estimados para su construcción, es decir, consideraremos que un modelo es mejor que otro si tiene un mejor resultado de MCC y tiene un número razonable de parámetros frente al resto de modelos.

4. Descripción de las Variables.

Antes de realizar ningún modelo Estadístico para predecir la probabilidad de que una empresa este cometiendo fraude es preciso "conocer" la información de la que partimos. Para ello se procede a continuación a realizar una breve descripción de cada una de las variables. La forma de operar será estudiar la evolución trimestral de cada una de las variables presentes. Para las variables de naturaleza cuantitativa se estudiará la evolución de su media y desviación típica, mientras que para las variables de naturaleza categórica se estudiará la evolución de la frecuencia de cada una de sus categorías a lo largo del tiempo.

Dado el origen '*camuflado*' de los datos, la interpretación de los estadísticos descriptivos no se puede realizar de forma clara. No obstante en el *Anexo I* podemos encontrar los gráficos correspondientes. En los gráficos que se muestran en dicho anexo para las variables cuantitativas se representa la evolución de la media y la desviación típica, de forma que en el eje horizontal están representados los 66 trimestres de los cuales se tiene información en el estudio, en el eje vertical izquierdo aparece representado la escala correspondiente a la media y en el eje vertical derecho se encuentra la escala de la desviación típica. Para las variables categóricas se muestra la evolución a lo largo de los 66 trimestres de los porcentajes observados para cada una de las categorías.

4.1. Variables Cuantitativas.

Se consideran variables cuantitativas todas aquellas que toman como argumento cantidades numéricas ya sean discretas o continuas.

4.1.1. Variables Socio-Económicas.

La familia de variables Socio-Económicas (**VarSocioEconomica1, VarSocioEconomica2, VarSocioEconomica3, VarSocioEconomica4, VarSocioEconomica5, VarSocioEconomica6, VarSocioEconomica7**) está formada por un total de siete variables medidas en escala de razón que miden siete aspectos diferentes sobre características sociales y económicas de las empresas registradas en la agencia tributaria.

Esta familia de variables tienen en común que los valores altos en dichas variables indican mala situación Socio-Económica, es decir, lo ideal para cada empresa es tener valores bajos en estas variables.

4.1.2. Variables relacionadas con la deuda.

Las cuatro siguientes variables pueden formar otra familia, la cual, recogen información sobre la deuda de cada empresa registrada en la agencia tributaria. Estas variables son:

Deuda: Cuantifica la deuda de cada empresa.

Deuda_Financiada: Cuantifica la deuda que ha sido financiada por el estado.

Num_Trimestres_En_Deuda: Cuantifica de forma discreta el número de trimestres en los que una empresa registrada en la agencia tributaria presenta una deuda con el Estado.

Pagado_de_la_Deuda: Cuantifica la cantidad que una empresa ha pagado de su deuda.

4.1.3. Variables relacionadas con el beneficio de las empresas.

Las variables que forman esta familia son dos:

Beneficio declarado: Cuantifica el beneficio que una empresa declara haber tenido en un trimestre determinado ante la agencia tributaria.

Beneficio declarado revisado: Cuantifica el beneficio que una empresa declara haber tenido en un trimestre determinado ante la agencia tributaria después de posibles subsanaciones que puedan haberse producido.

4.1.4. Variables relacionadas con el beneficio de las empresas por grupos.

Para cada uno de las empresas registradas en la Agencia Tributaria se realizó una agrupación en función de una serie características, utilizando como base esta agrupación se llevaron a cabo las siguientes variables:

Tot_estim_fraude_grup_trim: Para cada uno de los grupos anteriormente mencionados se calculó el fraude conjunto, lo cual, es lo que representa esta variable.

Porc_estim_fraude_grup_trim: Indica que porcentaje de fraude comete el grupo al que pertenece la empresa.

Porc_estim_fraude_grup_3años: Porcentaje de fraude armonizado en los 12 trimestres anteriores.

4.2. Variables Cualitativas.

Se consideran variables cualitativas a todas aquellas variables que toman como argumento una clasificación o modalidad. Cada modalidad que se presenta se denomina atributo o categoría y la medición consiste en una clasificación de dichos atributos.

4.2.1. Variables de Clasificación del CIF.

Esta familia está compuesta por un total de 10 variables (**clasificación_CIF1, …, clasificación_CIF9** y **sector**) que recogen información inherente a cada empresas (CIF).

4.2.2. Variables de alerta.

Esta familia está compuesta por un total de cinco variables (**Var_indicios_aut1, Var_indicios_aut2, Var_indicios_manuales1** y **Var_indicios_manuales2, departamento_inspeccionador**) que recogen información sobre la señal de alerta de fraude ya sea por mecanismos automáticos de la Agencia Tributaria o manuales producidos por agentes así como el departamento que emite dicha alerta.

4.2.3. Variables relacionadas con los estados de las inspecciones o las reclamaciones.

Esta familia está compuesta por un total de 5 variables (**Estado_orden_servicio1, Estado_orden_servicio2, Estado_reclamo, Medio_reclamo** y **Tipo_reclamo)** que recogen información inherente a los estados de las inspecciones o reclamaciones puestas a las empresas.

4.2.4. Variables inherentes a las anomalías una vez detectadas.

Esta familia está compuesta por un total de dos variables (**Irregularidad** y **Tipo_Irregularidad)** que recogen informacion acerca del tipo y estado de las irregularidades encontradas a cada empresa.

5. Variables Finales.

5.1. Introducción.

Una vez definidas y estudiadas las variables originales (información bruta) se decide que la mejor forma de trabajar es agrupando dicha información de alguna forma para conseguir que la información de la que se parte, información histórica, pase a ser información no histórica, es decir, debemos eliminar la temporalidad de los datos sin perder información (estas agrupaciones o transformaciones se llevan a cabo intentando maximizar la correlación de cada una de las variables con la variable respuesta). Es necesario debido a la gran cantidad de información de partida, en la cual se necesita recoger información acumulada en algunas variables y en otras, se precisa de solventar el problema de la multicolinealidad que impide realizar cualquier técnica estadística que nos planteemos.

5.2. Variable Socio-Económica.

Dada la naturaleza de la familia de variables Socio-Económicas y el similar comportamiento observado en el estudio descriptivo de las mismas (*Anexo I*) se intuye que dichas variables esconden una fuerte multicolinealidad, lo cual, produce un contexto idóneo para la realización de un análisis factorial con el propósito de disminuir la dimensión de dichas variables. Concretamente, se trata de encontrar un conjunto de $k<7$ factores que expliquen suficientemente las variables observadas perdiendo el mínimo de información.

El análisis factorial se llevara a cabo mediante el software estadístico SAS partiendo directamente de las variables Socio-Económicas observables y utilizando la suma media de cuadrados (SMC) como método de obtención de los factores.

A continuación, se muestra un resumen de la proporción de variabilidad que se explica con cada uno de los siete posibles factores a considerar en el análisis factorial.

Eigenvalues of the Reduced Correlation Matrix: Total = 6.43022407 Average = 0.91860344				
	Eigenvalue	Difference	Proportion	Cumulative
1	5.76880626	5.42201099	0.8971	0.8971
2	0.34679527	0.12546309	0.0539	0.9511
3	0.22133217	0.12072056	0.0344	0.9855
4	0.10061161	0.09048526	0.0156	1.0011
5	0.01012636	0.01857480	0.0016	1.0027
6	-.00844845	0.00055070	-0.0013	1.0014
7	-.00899915		-0.0014	1.0000

Tabla 5.2.1

Queda claro que el número de factores no observables a tener en cuenta para resumir la información de las siete variables Socio-Económicas es de uno, ya que con un único factor se consigue explicar casi el 90% (89.71%) de la información recogida en esta familia de variables.

Siendo la cantidad de variabilidad explicada de cada variable por el factor la siguiente:

Finales estimaciones comunalidad: Total = 5.768806						
Var_socioEconomica1	Var_socioEconomica2	Var_socioEconomica3	Var_socioEconomica4	Var_socioEconomica5	Var_socioEconomica6	Var_socioEconomica7
0.65155225	0.66337385	0.82575959	0.87092743	0.80638376	0.95824347	0.99256590

Tabla 5.2.2

La variable menos explicada por el factor es la primera (Var_socioeconomica1) con un 65.15% de variabilidad explicada, en el extremo opuesto se encuentra la séptima variable socioeconómica (Var_socioeconomica7) que es la mejor explicada por el factor con un 99.25% de variabilidad explicada.

A continuación se muestran las puntuaciones de cada una de las variables con el factor no observable:

Patrón Factor		
		Factor1
Var_socioEconomica1	Var_socioEconomica1	0.80719
Var_socioEconomica2	Var_socioEconomica2	0.81448
Var_socioEconomica3	Var_socioEconomica3	0.90871
Var_socioEconomica4	Var_socioEconomica4	0.93323
Var_socioEconomica5	Var_socioEconomica5	0.89799
Var_socioEconomica6	Var_socioEconomica6	0.97890
Var_socioEconomica7	Var_socioEconomica7	0.99628

Tabla 5.2.3

Todas las puntuaciones son positivas, es decir, todas las variables Socio-Económicas del estudio tienen entre sí (como ya se intuía) y entre el factor no observable una relación directa, es decir, a mayor valor de cada una de estas siete variables se espera un mayor valor en el factor.

Se presenta a continuación una tabla que recoge el estadístico KMO y la significación en el contraste de Bartlett:

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		,693
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	2,827E8
	gl	21
	Sig.	,000

Tabla 5.2.4

Ante los resultados anteriores podemos concluir que si es razonable realizar un análisis factorial para disminuir la dimensión de las variables socio-económicas dado que el estadístico KMO supera el valor 0.5 valiendo prácticamente 0.7 y los datos muestran evidencia estadística bajo cualquier nivel de significación razonable para rechazar la hipótesis nula de que la matriz de correlaciones entre las variables sea la matriz identidad.

Se concluye con esto que el factor no observable resultado del análisis multivariante factorial llevado a cabo será utilizado como una variable final apta para entrar en cualquiera de los modelos de predicción que se llevaran a cabo para predecir el fraude empresarial en España.

5.3. Variables del usuario.

A continuación se describen las variables que junto con el factor no observable que recoge la información de las variables Socio Económicas entrarán en los modelos predictivos a priori.

Variable respuesta (**Fraude**): Describe si una empresa será o no fraudulenta. Esta variable tomará tres valores. El valor "1" cuando se ha encontrado fraude en la empresa; valor "-1" cuando no se ha encontrado fraude en la empresa y valor "0" si no se han realizado visitas a esa empresa.

F1: Esta variable describe si una empresa ha sido o no fraudulenta por parte del sector1. Toma los mismos tres valores que la variable respuesta. El valor "1" cuando se ha encontrado fraude por parte del secto1en la empresa; valor "-1" cuando no se ha encontrado fraude en la empresa por parte del sector1 y el valor "0" si no se han realizado visitas por parte del sector1a esa empresa.

F2: Esta variable describe si una empresa ha sido o no fraudulenta por parte del sector2. Toma los mismos tres valores que la variable respuesta.

F3: Esta variable describe si una empresa ha sido o no fraudulenta por parte del sector3. Toma los mismos tres valores que la variable respuesta.

F4: Esta variable describe si una empresa ha sido o no fraudulenta por parte del sector4 servicio. Toma los mismos tres valores que la variable respuesta.

F6: Esta variable describe si una empresa ha sido o no fraudulenta por parte del sector6. Toma los mismos tres valores que la variable respuesta.

F7: Esta variable describe si una empresa ha sido o no fraudulenta por parte del sector7. Toma los mismos tres valores que la variable respuesta.

F8: Esta variable describe si una empresa ha sido o no fraudulenta por parte del sector8. Toma los mismos tres valores que la variable respuesta.

FI1: Esta variable genera un índice que pondera con un mayor peso a los fraudes correspondientes a la variable F1, realizados en los trimestres más cercanos al trimestre en estudio. Es importante tener en cuenta si la última vez que se la inspeccionó se le detectó fraude o no.

FI2, FI3, FI4, FI6, FI7, FI8: Todas estas variables se calculan de igual forma que FI1, pero teniendo en cuenta su variable correspondiente F2, F3, etc.

Clasificacion_CIF6_v2: Esta variable indica si una empresa pertenece o no a una zona especial. Toma el valor 1 si la empresa pertenece a una zona especial y el valor cero si el valor correspondiente a la variable que indica si la empresa pertenece a una zona especial toma valor perdido.

Sector_v2: Esta variable describe el fraude relativo, cociente entre el número de fraudes y el número de visitas de las empresas de la categoría correspondiente.

Clasificacion_CIF5_v2: Esta variable indica la zona a la que pertenece la empresa. Tendrá dos categorías: Zona1 y Zona2, al igual que la variable Clasificacion_CIF5.

Media_Clasificacion_CIF1: Esta variable describe el fraude relativo, cociente entre el número de fraudes y el número de visitas de las empresas de la categoría correspondiente: 'Categoria1', 'Categoria2' y 'NULL'. La categoría 'NULL' englobará las categorías 'Otros' y los datos perdidos de la variable Clasificacion_CIF1.

Tener_estimado_fraude_grupo: Tomará el valor 1 si la empresa tiene un fraude estimado para su grupo y cero en caso contrario.

Num_trimestres_en_deuda: Esta variable determina el número de trimestres en los que la empresa lleva en deuda.

Estm_fraude_grup_trim_v2: Esta variable determina las pérdidas trimestrales asociados a un grupo. Se pretende determinar si una empresa es más o menos fraudulenta. Esta variable toma tres valores: -1, 0 y 1. Para asignar estos valores se ha realizado una agregación de los datos, de modo que se han calculado los percentiles 25 y 75. Si la pérdida es inferior al percentil 25 la variable toma el valor -1, se le considerará menos fraudulenta. Si el valor de la pérdida se encuentra entre el percentil 25 y el 75 o NULL, a la variable Estm_fraude_grup_trim_v2 se le asigna el valor 0. Si el valor de la pérdida es superior al percentil 75 se le asigna a la variable el valor 1, por lo que a esa empresa se la considerará más fraudulenta.

Prob_local: Esta variable determina la probabilidad de fraude asignada a cada municipio.

Ant: Esta variable indicará el número de trimestres que lleva cada empresa registrada ante la Agencia Tributaria.

Trimestre: Tomará un valor de 1 al 66, siendo el 1 el primer trimestre del año 1997 y 66 el segundo trimestre del año 2014.

MP_Claficiacion_CIF7: Esta variable determina la media ponderada de los 10 últimos trimestres de la séptima variable de clasificación del CIF de cada empresa. La media corresponde al fraude relativo de la variable respuesta para cada una de las categorías de la séptima variable de clasificación del CIF agrupada. Si no existe una empresa en alguno de los últimos 10 trimestres se le asignará como media, la correspondiente a la categoría de la variable que contenga un mayor número de datos.

De forma similar se definirán las siguientes 10 variables, con la única salvedad de cómo se definen las categorías.

MP_Clasificacion_CIF8: Esta variable determina la media ponderada de los 10 últimos trimestres de la octava variable de clasificación del CIF de cada empresa. La media corresponde al fraude relativo de la variable respuesta para cada una de las categorías de la octava variable clasificadora del CIF agrupada. Si no existe una empresa en alguno de los últimos 10 trimestres se le asignará como media, la correspondiente a la categoría de la variable que contenga un mayor número de datos.

MP_Clasificacion_CIF3: Esta variable determina la media ponderada de los 10 últimos trimestres de la tercera variable de clasificación del CIF de cada empresa. La media corresponde al fraude relativo de la variable respuesta para cada una de las categorías de la tercera variable de clasificación del CIF agrupada. Si no existe una empresa en alguno de los últimos 10 trimestres se le asignará como media, la correspondiente a la categoría de la variable que contenga un mayor número de datos.

MP_var_indicios_manuales2: Esta variable determina la media ponderada de los 10 últimos trimestres de los indicios manuales2 de anomalía agrupado de cada empresa. La media corresponde al fraude relativo de la variable respuesta para cada una de las categorías de la variable var_indicios_manuales2. Si no existe una empresa en alguno de los últimos 10 trimestres se le asignará como media, la correspondiente a la categoría de la variable que contenga un mayor número de datos.

MP_estado_incidencia: Esta variable determina la media ponderada de los 10 últimos trimestres del estado de refacturación agrupado de cada empresa. La media corresponde al fraude relativo de la variable respuesta para cada una de las categorías de la variable estado de refacturación. Si no existe una empresa en alguno de los últimos 10trimestres se le asignará como media, la correspondiente a la categoría de la variable que contenga un mayor número de datos.

MP_medio_incidencia: Esta variable determina la media ponderada de los 10 últimos trimestres del medio de refacturación agrupado de cada usuario. La media corresponde al fraude relativo de la variable respuesta para cada una de las categorías de la variable medio de refacturación. Si no existe una empresa en alguno de los últimos 10 trimestres

se le asignará como media, la correspondiente a la categoría de la variable que contenga un mayor número de datos.

MP_var_indicios_aut2: Esta variable determina la media ponderada de los 10 últimos trimestres de los indicios automaticos2 de anomalía agrupado de cada empresa. La media corresponde al fraude relativo de la variable respuesta para cada una de las categorías de la variable medio de refacturación agrupado. Si no existe una empresa en alguno de los últimos 10 trimestres se le asignará como media, la correspondiente a la categoría de la variable que contenga un mayor número de datos.

MP_var_indicios_aut4: Esta variable determina la media ponderada de los 10 últimos trimestres de los indicios automaticos4 de anomalía agrupado de cada empresa. La media corresponde al fraude relativo de la variable respuesta para cada una de las categorías de la variable medio de refacturación agrupado. Si no existe una empresa en alguno de los últimos 10 trimestres se le asignará como media, la correspondiente a la categoría de la variable que contenga un mayor número de datos.

MP_ var_indicios_manuales1: Esta variable determina la media ponderada de los 10 últimos trimestres de los indicios manuales1 de anomalía agrupado de cada empresa. La media corresponde al fraude relativo de la variable respuesta para cada una de las categorías de la variable var_indicios_manuales2 agrupada. Si no existe una empresa en alguno de los últimos 10 trimestres se le asignará como media, la correspondiente a la categoría de la variable que contenga un mayor número de datos.

MP_ var_indicios_aut3: Esta variable determina la media ponderada de los 10 últimos trimestres de los indicios automaticos3 de anomalía agrupado de cada empresa. La media corresponde al fraude relativo de la variable respuesta para cada una de las categorías de la variable medio de refacturación agrupado. Si no existe una empresa en alguno de los últimos 10 trimestres se le asignará como media, la correspondiente a la categoría de la variable que contenga un mayor número de datos.

MP_ Clasificacion_CIF9: Esta variable determina la media ponderada de los 10 últimos trimestres de la novena variable de clasificación del CIF de cada empresa. La media corresponde al fraude relativo de la variable respuesta para cada una de las categorías de la novena variable de clasificación del CIF agrupada. Si no existe una empresa en alguno de los últimos 10 trimestres se le asignará como media, la correspondiente a la categoría de la variable que contenga un mayor número de datos.

Con la misma filosofía, pero sin computar su media, definiremos la siguiente variable:

MP_definitivaVSprovisional: Esta variable determina la media ponderada de los 10 últimos trimestres de definitivaVSprovisional de cada empresa, que ha sido codificada

previamente. Cuando no exista una empresa en alguno de los trimestres anteriores se le pondrá el valor 0.

Por último, se definen las variables relativas al consumo de las empresas.

Indice_beneficio: Esta primera variable se determina mediante un cociente, donde en el numerador está la diferencia entre el beneficio de una empresa y el beneficio medio de las empresas en ese trimestre que pertenecen a su mismo grupo. Y en el denominador este mismo beneficio medio.

Media_beneficio: Es el valor medio de las anteriores diferencias para una empresa en los últimos 10 trimestres.

Variabilidad_beneficio: Es la desviación típica de las anteriores diferencias para una empresa en los últimos 10 trimestres.

MP_dif_beneficio: Esta variable determina la variación en el beneficio de una empresa respecto al valor medio. Para ello se ha definido la variable dif_beneficio, que tomará valor 1 si la diferencia es negativa, 0 si la diferencia es cero y -1 si la diferencia es positiva. Posteriormente se realizará una media ponderada de los últimos 10 trimestres. Cuando no exista una empresa en alguno de los trimestres anteriores se le asignará a la variable el valor 0.

6. Modelos de Predicción.

6.1. Introducción.

Una vez alcanzado este punto llegamos al punto donde afrontaremos de forma directa el principal objetivo del presente estudio, el cual, siempre ha sido predecir de una u otra forma la probabilidad con que cada una de las empresas registradas ante la Agencia Tributaria puede o no estar cometiendo fraude al Estado declarando un beneficio menor al real.

Para abordar este objetivo se utilizarán 5 técnicas estadísticas de predicción, la Regresión Logística, Redes Neuronales, Random Forest, Gradient Boosting y Ensamblado de Modelos; para estos métodos se proporcionarán medidas de ajuste y comparativas entre ambos para decidir cuál de los modelos conseguidos es el que mejor resuelve nuestro objetivo.

6.2. Fragmentación de la información.

Se decide que la mejor forma de llevar a cabo las diferentes técnicas de predicción, así, como su posterior validación y pruebas, será dividiendo el total de datos o información en tres archivos distintos, donde cada uno de ellos contará con la información relativa a determinados trimestres sucesivos. La división es la siguiente:

Aprendizaje: Este conjunto de datos estará formado por la información relativa a los 47 primeros trimestres, seleccionando únicamente a aquellas empresas a las que se tenga constancia de haberse realizado alguna inspección en los últimos trimestres y con una antigüedad considerable ante la Agencia Tributaria. Dicha información será utilizada por cada uno de los modelos de predicción que se van a llevar a cabo para predecir los valores de los parámetros inherentes a cada modelo.

Validación: Este conjunto de datos estará formado por la información relativa a los 12 trimestres posteriores a los utilizados en el conjunto de aprendizaje, es decir, a los trimestres que van desde el 48 al 59 ambos inclusive, seleccionando únicamente a aquellas empresas a las que se tenga constancia de haberse realizado alguna inspección en los últimos trimestres y con una antigüedad considerable ante la Agencia Tributaria. Dicha información será utilizada para decidir cuál de todos los modelos de predicción llevados a cabo es el mejor, es decir, cuál de ellos comete un menor error en la estimación.

Test: Este conjunto de datos estará formado por la información relativa a los 7 últimos trimestres, es decir, a los trimestres que van del 60 al 66 ambos inclusive, seleccionando únicamente a aquellas empresas a las que se tenga constancia de haberse realizado alguna inspección en los últimos trimestres y con una antigüedad considerable ante la Agencia Tributaria. Dicha información será utilizada para predecir de forma insesgada el grado de acierto de los modelos seleccionados como mejores en validación.

6.3. Regresión Logística.

6.3.1. Construcción del modelo.

La construcción del mejor modelo de Regresión Logística se lleva a cabo de forma empírica en un proceso de prueba-error donde se crea una parrilla de valores para los diferentes parámetros propios de esta técnica. Calculándose el coeficiente de correlación Matthews para cada uno de los modelos construidos.

La parrilla de valores para los diferentes parámetros son:

- Método de selección de variables (Stepwise, Backward o Forward).
- Probabilidad de entrada de una variable (`0.1, 0.05, 0.01, 0.001, 0.0001, 0.00001 y 0.000001`). Este parámetro no tiene sentido en el método de selección de variables Backward.
- Probabilidad de salida de una variable (`0.1, 0.05, 0.01, 0.001, 0.0001, 0.00001 y 0.000001`). Este parámetro no tiene sentido en el método de selección de variables Forward.
- Función de enlace (`Probit, Logit o Cloglog`)[8].
- Puntos de corte. El modelo de Regresión Logística clasificará como usuario fraudulento a aquellos que tengan una probabilidad de cometer fraude entre un valor que denominaros "pmin" y "pmax", estos dos valores no son parámetros inherentes a la Regresión Logistica, sino que se han calculado a partir de los datos de validación para evaluar el comportamiento o la efectividad de los modelos. La parrilla de valores para estos dos puntos de corte son:

    - Pmin tomará todos los valores entre 0.2 y 0.7 equidistantes 0.1.
    - Pmax tomará todos los valores entre 0.8 y 1.0 equidistantes 0.1.

En definitiva se construyen un total de 1323 modelos de Regresión Logística resultado de todas las posibles combinaciones de la parrilla de valores.

Se muestra a continuación la cabecera de una tabla resumen para comparar los diferentes modelos de Regresión Logísitca creados.

	Metodo	Funcion	VN	FP	FN	VP	num_parametros	Matthews	p_entrada	p_salida	pmax	pmin
1	FORWARD	Cloglog	76988	27436	41220	75438	29	0.3842600453	0.01	0	1	0.4
2	FORWARD	Cloglog	76978	27446	41215	75443	31	0.3842034658	0.1	0	1	0.4
3	FORWARD	Cloglog	76978	27446	41215	75443	31	0.3842034658	0.05	0	1	0.4
4	BACKWARD	Cloglog	76978	27446	41215	75443	31	0.3842034658	0	0.1	1	0.4
5	BACKWARD	Cloglog	76978	27446	41215	75443	31	0.3842034658	0	0.05	1	0.4
6	BACKWARD	Cloglog	76978	27446	41215	75443	31	0.3842034658	0	0.01	1	0.4
7	BACKWARD	Cloglog	76978	27446	41215	75443	31	0.3842034658	0	0.001	1	0.4
8	BACKWARD	Cloglog	76978	27446	41215	75443	31	0.3842034658	0	0.0001	1	0.4
9	BACKWARD	Cloglog	76978	27446	41215	75443	31	0.3842034658	0	0.00001	1	0.4
10	BACKWARD	Cloglog	76978	27446	41215	75443	31	0.3842034658	0	1E-6	1	0.4

Tabla 6.3.1.1 Resumen Modelos Logística

Se obtiene que el valor máximo en el coeficiente de correlación de Matthews (0.3842) es conseguido por el modelo de Regresión Logística con los siguientes parámetros:

- Método de selección de variables: Forward.
- Probabilidad de entrada de una variable: 0.01
- Probabilidad de salida de una variable: No procede.
- Función de enlace: Cloglog.
- Puntos de corte:
  - Pmin: 0.4.
  - Pmax: 1.

Dicho modelo de Regresión logística consta de 29 parámetros a estimar, lo cual, es un número bastante atractivo dado que si quisiéramos reducirlo deberíamos disminuir mucho el coeficiente de Matthew.

Una vez encontrada la mejor estructura para nuestro modelo de Regresión Logística se procede a la construcción de un modelo que replique dicha estructura y que cuente con todas las posibles interacciones de orden 2 en la selección de variables. El motivo por el cual no se han tenido en cuenta las interacciones como fuentes de variación en los modelos probados es por limitaciones temporales inherentes al hardware utilizado.

El modelo construido replicando la mejor estructura probada y teniendo en cuenta todas las posibles interacciones de orden 2 resulta ser un modelo peor que el conseguido sin las interacciones. Los motivos de este suceso es que el modelo que cuenta con interacciones termina estimando un total de 244 parámetros y obteniendo un coeficiente

de Matthew ligeramente inferior al mismo modelo con solo los efectos principales. Esto se debe a que se está produciendo una clara sobreparametrización sobre los datos de validación, por ello se considerara que el mejor modelo de Regresión Logística es el comentado a partir de la tabla anterior 6.3.1.1. En el anexo correspondiente a Regresión Logística (Punto 9.2.) se muestran más características del modelo con interacciones probado.

A continuación, se muestra una tabla resumen que muestra las variables que han sido consideradas significativamente relevantes en el modelo de regresión logística, cada variable aparecerá acompañada por su nivel de significación en el modelo, su parámetro estimado correspondiente, el error típico de dicho parámetro y su respectivo intervalo de confianza.

Parameter	Wald Chi-Square	Pr > ChiSq	Estimate	Standard Error	95% Confidence Limits	
Intercept	42.2564	<.0001	-5.5968	0.8610	-7.2843	-3.9093
CLASIFICACION_CIF6_v2 (0)	365.8825	<.0001	0.0596	0.00312	0.0535	0.0657
CLASIFICACION_CIF5_v2 (0)	42.2355	<.0001	-0.0235	0.00362	-0.0306	-0.0164
Tener_estimado_fraude_grupo (0)	4336.3812	<.0001	0.2242	0.00341	0.2176	0.2309
estm_fraude_grup_trim_v2 (-1)	270.8915	<.0001	0.1066	0.00648	0.0939	0.1193
estm_fraude_grup_trim_v2 (0)	0.5225	0.4698	-0.00354	0.00490	-0.0132	0.00607
F1	143.6471	<.0001	-0.0316	0.00264	-0.0367	-0.0264
F4	48.3712	<.0001	0.1425	0.0205	0.1023	0.1826
F7	1070.2283	<.0001	0.1322	0.00404	0.1243	0.1401
F8	11.8708	0.0006	-0.3577	0.1038	-0.5612	-0.1542
Factor	536.7032	<.0001	0.0645	0.00279	0.0591	0.0700
MP_CLASIFICACION_CIF3	52299.4451	<.0001	3.7184	0.0163	3.6866	3.7503
MP_CLASIFICACION_CIF7	1166.2371	<.0001	1.3990	0.0410	1.3187	1.4793
MP_CLASIFICACION_CIF8	6586.0635	<.0001	7.5578	0.0931	7.3753	7.7403
MP_CLASIFICACION_CIF9	181.0451	<.0001	-1.0193	0.0758	-1.1677	-0.8708
MP_DefinitivaVSProvisional	2488.2440	<.0001	-0.2411	0.00483	-0.2506	-0.2316
MP_Var_indicios_aut2	22.1106	<.0001	1.0366	0.2204	0.6045	1.4686
MP_Var_indicios_aut4	15.0224	0.0001	-1.0360	0.2673	-1.5599	-0.5121
MP_estado_reclamo_refact	6.6981	0.0097	-6.1605	2.3803	-10.8259	-1.4951
Num_trimestres_en_deuda	35.4229	<.0001	0.000894	0.000150	0.000600	0.00119
antiguedad	703.4070	<.0001	0.0894	0.00337	0.0828	0.0960
indice_beneficio	142.2247	<.0001	-0.0514	0.00431	-0.0599	-0.0430
media_beneficio	229.7506	<.0001	0.0810	0.00535	0.0705	0.0915
prob_local	9793.8177	<.0001	1.9463	0.0197	1.9077	1.9848
sector_v2	448.9843	<.0001	0.5561	0.0262	0.5047	0.6076
trimestre	3591.7962	<.0001	0.0136	0.000227	0.0132	0.0140
variabilidad_beneficio	130.8065	<.0001	0.0917	0.00802	0.0760	0.1074

Tabla 6.3.1.2

De 32 variables regresoras introducidas en el modelo de regresión logística, éste ha seleccionado como significativas a 26, todas ellas con un p-valor inferior a 0.001 salvo la variable estm_fraude_grup_trim_v2 (0) (p-valor=0.4698), es decir, para una confianza de alpha=5% los datos muestran evidencia estadística para rechazar la siguiente hipótesis nula:

H_0: $\beta i=0$ frente a H_1: $\beta i\neq 0$

Por lo que las 26 variables anteriormente expuestas son consideradas relevantes a la hora de predecir la probabilidad de fraude. Cabe destacar que la variable dummy que representa a la categoría cero (nivel medio de fraude estimado por grupo) de la variable estm_fraude_grup_trim_v2 no se ha considerado significativamente distinta a su categoría de referencia (nivel alto de fraude estimado por grupo) en el análisis.

6.3.2. Interpretación del modelo.

En la siguiente tabla se recogen los OR estimados y su intervalo de confianza al 95% para cada variable. Los OR como ya se describieron en la metodología miden el cambio de probabilidad que se produce por el cambio en la categoría de cada variable respecto a la categoría marcada como referencia en el caso de variables categóricas o dicotómicas. En el caso de variables continuas, el OR es una media del cambio en la probabilidad logística al aumentar una unidad la variable predictiva, es decir nos da una tendencia general del comportamiento de la variable respuesta frente a la variable del modelo.

Odds Ratio Estimates and Wald Confidence Intervals				Descriptivos			
Effect	Estimate	95% Confidence Limits		Mínimo	Máximo	Q1	Q3
CLASIFICACION_CIF6_v2 (0)	1,205000	1,185000	1,225	.	.	.	.
CLASIFICACION_CIF5_v2 (0)	0,840000	0,825000	0,855	.	.	.	.
Tener_estimado_fraude_grupo (0)	1,769000	1,739000	1,8	.	.	.	.
estm_fraude_grup_trim_v2 (-1)	1,312000	1,278000	1,346	.	.	.	.
estm_fraude_grup_trim_v2 (0)	1,151000	1,130000	1,173	-9,330273	6,900000	-0,0015625	0,000000
F1	0,960000	0,953000	0,966	-9.3302734	6.9000000	-0.0015625	0
F4	1,240000	1,176000	1,309	-5.3000000	2.0000000	0	0
F7	1,255000	1,240000	1,27	-5.1625000	4.4015625	0	0
F8	0,747000	0,575000	0,972	-1.0000000	0.9000000	0	0
Factor	1,205000	1,185000	1,225	-0.5887443	2.8323752	-0.5887443	-0.2332122
MP_CLASIFICACION_CIF3	196,204	187,083	205,769	0.0851124	0.6502133	0.2787127	0.3873542
MP_CLASIFICACION_CIF7	6,862000	6,112000	7,704	0.3168062	0.6716342	0.3168062	0.4166943
MP_CLASIFICACION_CIF8	>999.999	>999.999	>999.999	0.0983444	0.3709605	0.3709605	0.3709605
MP_CLASIFICACION_CIF9	0,353000	0,288000	0,432	0.1211499	0.5227952	0.3501125	0.3501125
MP_DefinitivaVSProvisional	0,745000	0,726000	0,764	-1.0000000	1.0000000	-1.0000000	-0.7264285
MP_Var_indicios_aut2	8,485000	4,065000	17,709	0.3541817	0.6654528	0.3541817	0.3541817
MP_Var_indicios_aut4	0,034000	0,014000	0,082	0.3541921	0.6012912	0.3541921	0.3541921
MP_estado_reclamo_refact	<0.001	<0.001	0,008	0.3607821	0.4252623	0.3607821	0.3607821
Num_trimestres_en_deuda	1,001000	1,001000	1,001	0	155.0000000	0	6.0000000
antiguedad	1,088000	1,079000	1,097	8.0000000	12.0000000	12.0000000	12.0000000
indice_beneficio	1,123000	1,107000	1,14	-1.0000000	63.7843246	-0.6304462	0.1723974
media_beneficio	8,485000	4,065000	17,709	0,354182	0,665453	0,3541817	0,354182
prob_local	20,613000	19,512000	21,777	-1.0000000	55.9173909	-0.5160158	0.1666965
sector_v2	1,894000	1,770000	2,028	0.1693254	0.4794661	0.2774918	0.4794661
trimestre	1,020000	1,019000	1,02	8.0000000	47.0000000	15.0000000	38.0000000
variabilidad_beneficio	1,215000	1,188000	1,243	0	15.9093431	0.1022765	0.3581175

Tabla 6.3.3.1

Cabe destacar que hay bastantes variables continuas cuya amplitud o recorrido es inferior a la unidad, como por ejemplo pasa con todas las variables no estáticas (excepto '*MP_DefinitivaVSprovisional'*) por lo que la estimación del OR puede ser desmesurada.

No obstante a partir de la tabla anterior podemos sacar numerosas conclusiones, algunas de las cuales son por ejemplo:

Tener_estimado_fraude_grupo (0): La probabilidad de ser una empresa fraudulenta no teniendo un fraude estimado para el grupo al que pertenece aumenta en 0.769 respecto a las empresas que si tienen un fraude estimado por grupo.

Clasificacion_CIF5_v2 (0): La probabilidad de ser una empresa fraudulenta estando en la zona1 es 0.84 veces la probabilidad de ser una empresa fraudulenta estando en la zona2.

Sector_v2: Por cada unidad que aumenta el fraude relativo la probabilidad de ser una empresa fraudulenta lo hace en 0.894.

Trimestre: Por cada trimestre que pasa se espera en media un incremento en la probabilidad de fraude en una empresa de 0.020, es decir, por cada trimestre que transcurre las empresas en media son un 2% más fraudulentas.

Ningún intervalo de confianza incluye la unidad, lo cual, es lo deseable, porque en el caso de que no fuera así indicaría que para el nivel de confianza alpha=0,05 utilizado no se podría garantizar que para el cambio en una unidad o categoría de una variable independiente produjera un cambio en la probabilidad estimada de fraude.

6.3.3. Conclusión.

Se ha llevado a cabo una exhaustiva búsqueda de la mejor estructura de Regresión Logística, en el que se probaron un total de 1323 modelos, resultado de combinar un conjunto de parámetros con diferente parrilla de valores. La estimación del modelo se llevó a cabo a partir de los datos de Entrenamiento o Aprendizaje, mientras que la comparación o la evaluación de dichos modelos se realizaron sobre el conjunto de datos de Validación.

Por último y para concluir con este este apartado se va a proporcionar una estimación insesgada del grado de acierto de nuestro modelo óptimo, para lo cual, se utilizará el conjunto de datos Test.

Se muestra a continuación una tabla de contingencia entre los valores observados y predichos por el modelo para la variable dependiente Fraude:

		Fraude Estimado	
		-1	1
Fraude	-1	29399	6617
	1	12776	15290

- La probabilidad de clasificar de forma correcta a una empresa cualquiera registrada ante la Agencia Tributaria es del 0,69.
- La probabilidad de clasificar de forma correcta a una empresa no fraudulenta registrada ante la Agencia Tributaria es del 0,81 (especificidad).
- La probabilidad de clasificar de forma correcta a una empresa fraudulenta registrada ante la Agencia Tributaria es del 0,54 (sensibilidad).

6.4. Redes Neuronales.

6.4.1. Construcción de la red.

La construcción de la mejor Red Neuronal se lleva a cabo de forma empírica en un proceso de prueba-error donde se crea una parrilla de valores para los diferentes parámetros propios de esta técnica. Calculándose el coeficiente de correlación Matthews para cada una de las Redes construidas.

La parrilla de valores para los diferentes parámetros son:

- Función de activación (Tangente Hiperbólica, exponencial, Elliott o Arco tangente)[9].
- Número de nodos en la capa oculta (1,2,3,…,20). Los motivos por los que solo se utiliza una capa oculta y por los que a dicha capa se le limita el número de nodos a un máximo de 20 es por limitaciones temporales inherentes al hardware utilizado.
- Puntos de corte. Los modelos de Redes Neuronales clasificarán como usuarios fraudulentos a aquellos que tengan una probabilidad de cometer fraude entre un valor que denominaros “pmin” y “pmax”. La parrilla de valores para estos dos puntos de corte son:
  - Pmin tomará todos los valores entre 0.2 y 0.7 equidistantes 0.1.
  - Pmax tomará todos los valores entre 0.8 y 1.0 equidistantes 0.1.

En definitiva se construyen un total de 1440 modelos de Redes Neuronales resultado de todas las posibles combinaciones de la parrilla de valores.

Se muestra a continuación la cabecera de una tabla resumen para comparar los diferentes modelos de Redes Neuronales creados.

	f_activacion	VN	FP	FN	VP	Matthews	nodos1	pmax	pmin
1	TAN	72139	32285	35542	81116	0.385696849	17	1	0.5
2	TAN	79624	24800	44328	72330	0.3847803336	1	0.9	0.5
3	TAN	79624	24800	44328	72330	0.3847803336	1	1	0.5
4	TAN	81712	22712	46976	69682	0.3844547396	17	1	0.6
5	TAN	74489	29935	38763	77895	0.3805838131	20	1	0.5
6	TAN	80558	23866	46184	70474	0.3790766958	2	1	0.5
7	TAN	77358	27066	42320	74338	0.3787537316	3	1	0.4
8	TAN	77632	26792	42801	73857	0.3774751447	10	1	0.4
9	TAN	84137	20287	51324	65334	0.3748613475	20	1	0.6
10	TAN	77905	26519	43448	73210	0.3748306811	5	1	0.4

Tabla 6.4.1.1. Resumen Modelos Redes

Se obtiene que el valor máximo en el coeficiente de correlación de Matthews (0.3856) es conseguido por el modelo de Redes Neuronales con los siguientes parámetros:

- Función de activación: Tangente Hiperbólica
- Número de nodos en la capa oculta: 17
- Puntos de corte:
  - Pmin: 0.5
  - Pmax: 1.0

No obstante se observa en la tabla resumen anterior que reduciendo en menos de un 0.01 nuestro valor máximo del coeficiente de Matthews tenemos una estructura de Red Neuronal con tan solo un único nodo. Esto hace que dicha estructura alternativa sea mucho más llamativa en cuanto a la parametrización de la Red Neuronal. Calculamos ahora el número de parámetros a estimar en cada una de estas dos estructuras:

El número de parámetros a estimar por una Red Neuronal con una sola capa y una variable output es “h(k+1)+h+1”, donde “h” es el número de nodos ocultos y “k” el número de variables de entrada.

- Número de parámetros a estimar con 17 nodos en la capa oculta=596.
- Número de parámetros a estimar con 1 nodos en la capa oculta=36.

Una vez calculado el número de parámetros a estimar en cada uno de los dos modelos parece evidente considerar que la estructura con un solo nodo en la capa oculta es la mejor.

Se muestra a continuación los valores de los parámetros estimados para su construcción:

- Función de activación: Tangente Hiperbólica.
- Número de nodos en la capa oculta: 1.
- Puntos de corte:
  - Pmin: 0.5.
  - Pmax: 0.9.

Con el objetivo de evaluar la "importancia" de cada una de las variables independientes dentro de la Red Neuronal y dada la imposibilidad de conseguir dicho resultado mediante el procedimiento "Proc Neural" de SAS, se procede al cálculo de un total de 33 Redes Neuronales más, repitiendo la estructura de la Red seleccionada anteriormente y eliminando en cada una de las Redes una única variable explicativa. Posteriormente se calculan las diferencias entre los Coeficientes de correlación de Matthews de estas nuevas Redes y el mismo valor alcanzado con la Red Seleccionada (0.3847). Se obtiene que eliminando 11 variables (**F6**, **F8**, **MP_Var_indicios_aut3**, **MP_salto_beneficio**, **antigüedad**, **indice_beneficio**, **media_CLASIFICACION_CIF1**, **media_beneficio**, **sector_v2**, **variabilidad_beneficio**, **CLASIFICACION_CIF6_v2**) de forma independiente se mejora la Red, por lo que se calcula una nueva Red Neuronal eliminando las variables expuestas anteriormente y se comprueba que realizando la eliminación conjunta la Red resultante sigue mejorando en términos de Coeficiente de Correlación de Matthews y también en términos de complejidad (menor número de variables menor complejidad).

Se repite el proceso anterior un total de tres veces hasta que la Red creada deja de mejorar cuando se le descartan variables. Se alcanza una vez terminado este proceso una nueva Red Neuronal que mejora todas las Redes probadas anteriormente. Dicha Red repite la estructura que en este mismo apartado se consideró óptima y consta de un total de 18 variables explicativas (**F1**, **F4**, **F7**, **MP_CLASIFICACION_CIF3**, **MP_CLASIFICACION_CIF7**, **MP_CLASIFICACION_CIF8**, **MP_CLASIFICACION_CIF9**, **MP_DefinitivaVSProvisional**, **MP_Var_indicios_aut2**, **MP_Var_indicios_aut4**, **MP_Var_indicios_manuales1**, **MP_estado_reclamo_refact**, **MP_medio_reclamo_refact**, **MP_var_indicios_manuales2**, **Num_trimestres_en_deuda**, **prob_local**, **CLASIFICACION_CIF5_v2**, **estm_fraude_grup_trim_v2**). Esta nueva Red alcanza un valor de Matthews de 0.3921 y consta de un total de 21 parámetros a estimar (15 menos que la anterior Red óptima).

6.4.2. Interpretación del modelo.

Una vez alcanzada la Red Neuronal donde todas las variables participan de forma positiva, se procede a calcular la importancia de cada una de las variables explicativas. Dicha importancia se establecerá en términos de la variabilidad del Coeficiente de

Correlación de Matthews construyendo 18 nuevas Redes que excluyan una sola variable explicativa.

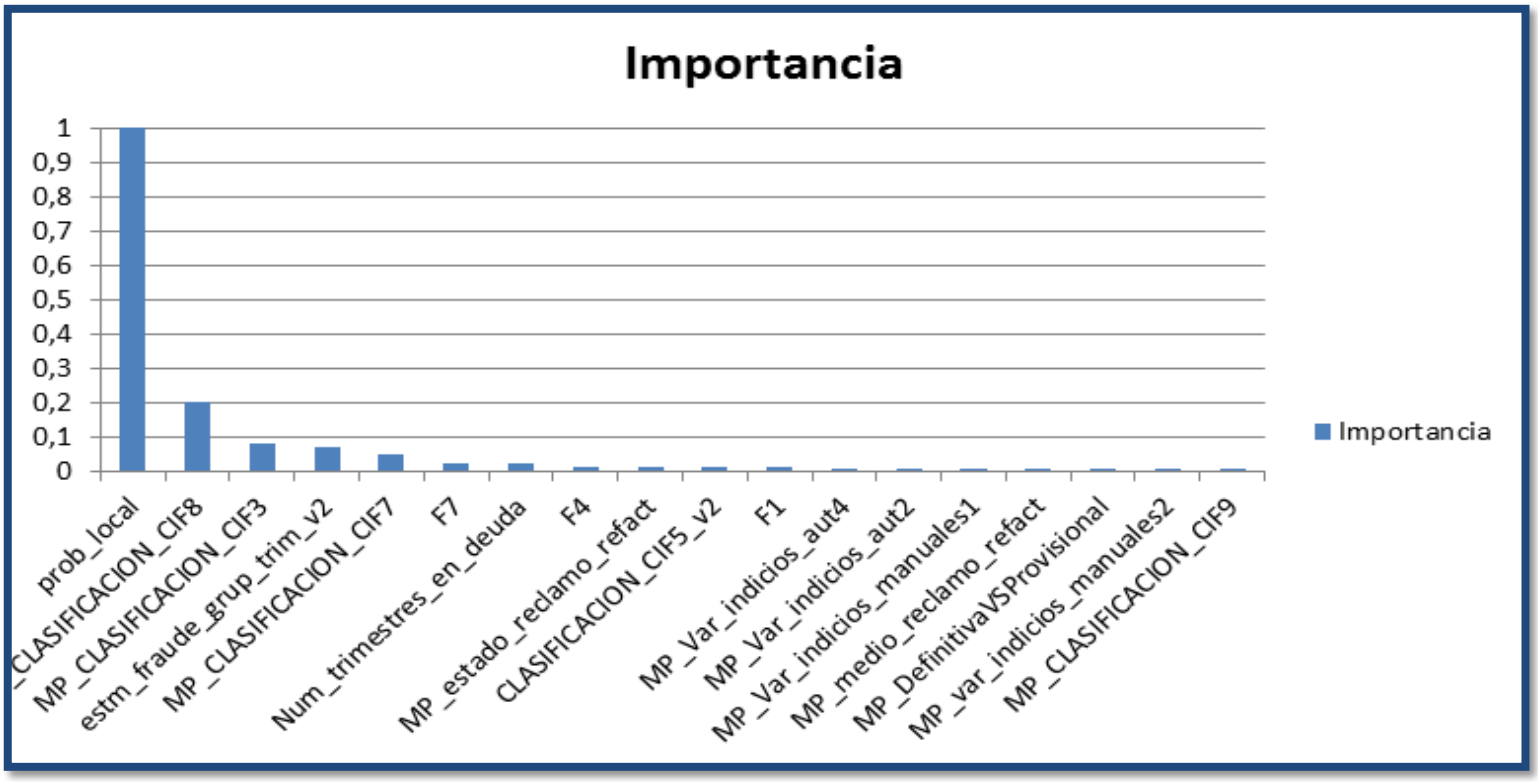


Gráfico 6.4.2.1 1. Importancia de las variables en el algoritmo de Redes Neuronales.

De esta forma se puede evaluar de forma rápida la importancia de las variables explicativas dentro de la Red Neuronal. Podemos apreciar que la variable con mayor importancia en la Red es 'Prob_local', la cual, determina la probabilidad de fraude asignada a cada municipio. El resto de variables aparentemente tiene una importancia relativa bastante inferior a dicha variable.

6.4.3. Conclusión.

Se ha llevado a cabo una exhaustiva búsqueda de la mejor estructura de Redes Neuronales, en el que se probaron un total de 1440 modelos, resultado de combinar un conjunto de parámetros con diferente parrilla de valores. Una vez fijada la mejor estructura se ha procedido a reducir la dimensión de las variables explicativas. La estimación del modelo se llevó a cabo a partir de los datos de Entrenamiento o Aprendizaje, mientras que la comparación o la evaluación de dichos modelos se realizaron sobre el conjunto de datos de Validación.

Por último y para concluir con este este apartado se va a proporcionar una estimación insesgada del grado de acierto de nuestro modelo óptimo, para lo cual, se utilizará el conjunto de datos Test. Se muestra a continuación una tabla de contingencia entre los valores observados y predichos por el modelo para la variable dependiente Fraude:

		Fraude Estimado	
		-1	1
Fraude	-1	28284	5706
	1	13891	16201

- La probabilidad de clasificar de forma correcta a una empresa cualquiera registrada ante la Agencia Tributaria es del 0,69.
- La probabilidad de clasificar de forma correcta a una empresa no fraudulenta registrada ante la Agencia Tributaria es del 0,83 (especificidad).
- La probabilidad de clasificar de forma correcta a una empresa fraudulenta registrada ante la Agencia Tributaria es del 0,53 (sensibilidad).

Gráfico 6.4.1.1

6.5. Random Forest.

6.5.1. Construcción del Random Forest.

La construcción del mejor Random Forest se lleva a cabo de forma empírica en un proceso de prueba-error donde se crea una parrilla de valores para los diferentes parámetros propios de esta técnica. Calculándose el coeficiente de correlación Matthews para cada una de las Redes construidas.

La parrilla de valores para los diferentes parámetros son:

- P-valor necesario para generar una regla de división (0.01, 0.05 y 0.1).
- Tamaño de hoja mínimo (50, 100, 600, 1100, 1600 y 2100).
- Máximo número de árboles a crear (50, 100 y 150).
- Numero de variables a tener en cuenta uno de los nodos de los diferentes arboles (5, 10, 15, 20 y 25).
- Porcentaje de la población que se muestrea en la construcción de cada árbol (valor que dejamos fijo en un 60%).
- Punto de corte. El modelo de Random Forest clasificará como usuario fraudulento a aquellos que tengan una probabilidad de cometer fraude superior al punto de corte, este valor no es un parámetro inherente a la construcción de nuestras estructuras de Random Forest, sino que se han calculado a partir de los datos de validación para evaluar el comportamiento o la efectividad de los modelos. El punto de corte tomará todos los valores entre 0.5 y 0.91 equidistantes 0.01.

En definitiva se construyen un total de 11070 modelos de Random Forest resultado de todas las posibles combinaciones de la parrilla de valores.

Se muestra a continuación la cabecera de una tabla resumen para comparar los diferentes modelos de Random Forest creados:

	VN	FP	FN	VP	Numrules	Matthews	PuntoDeCorteMin	maxtrees	leafsize	maxdepth	alpha	numvariables
1	88661	15763	56172	60486	97606	0.3860191951	0.53	100	100	25	0.05	10
2	87705	16719	54903	61755	97606	0.3852618943	0.52	100	100	25	0.05	10
3	88237	16187	55693	60965	109242	0.3850032729	0.53	100	100	25	0.05	15
4	88690	15734	56370	60288	32899	0.3848004931	0.52	50	100	25	0.05	5
5	86719	17705	53581	63077	97606	0.3847883801	0.51	100	100	25	0.05	10
6	87593	16831	54808	61850	32899	0.3847650849	0.51	50	100	25	0.05	5
7	89244	15180	57197	59461	109242	0.3846253383	0.54	100	100	25	0.05	15
8	87543	16881	54759	61899	65877	0.3845957027	0.51	100	100	25	0.05	5
9	87193	17231	54267	62391	109242	0.3845939414	0.52	100	100	25	0.05	15
10	89579	14845	57698	58960	97606	0.3845479502	0.54	100	100	25	0.05	10

Tabla 6.5.1.1. Resumen Modelos Random Forest.

Se obtiene que el valor máximo en el coeficiente de correlación de Matthews (0.3860) es conseguido por el modelo de Random Forest con la siguiente estructura:

- P-valor: 0.05.
- Tamaño de hoja mínimo: 100
- Máximo número de árboles a crear: 100.
- Numero de variables a tener en cuenta uno de los nodos de los diferentes arboles: 10.
- Porcentaje de la población que se muestrea en la construcción de cada árbol: 60%.
- Punto de corte: 0.53.

Cabe mencionar que los diferentes modelos creados aparentemente se muestran bastante estables ante las variaciones propuestas de los parámetros estructurales "Número máximo de árboles" y "P-valor".

6.5.2. Interpretación del modelo.

Importancia de las variables independientes: La importancia de una variable independiente es una medida que indica cuánto cambia el valor pronosticado por el modelo de Random Forest para diferentes valores de la variable independiente. La importancia normalizada es el resultado de los valores de importancia divididos por el valor de importancia mayor, expresados como porcentajes.

Se muestra a continuación un gráfico de barras de los valores de la tabla de importancia, clasificado en valor de importancia descendente. La tabla que recoge los valores de importancia e importancia normalizada se encuentra en el Anexo de Random Forest.

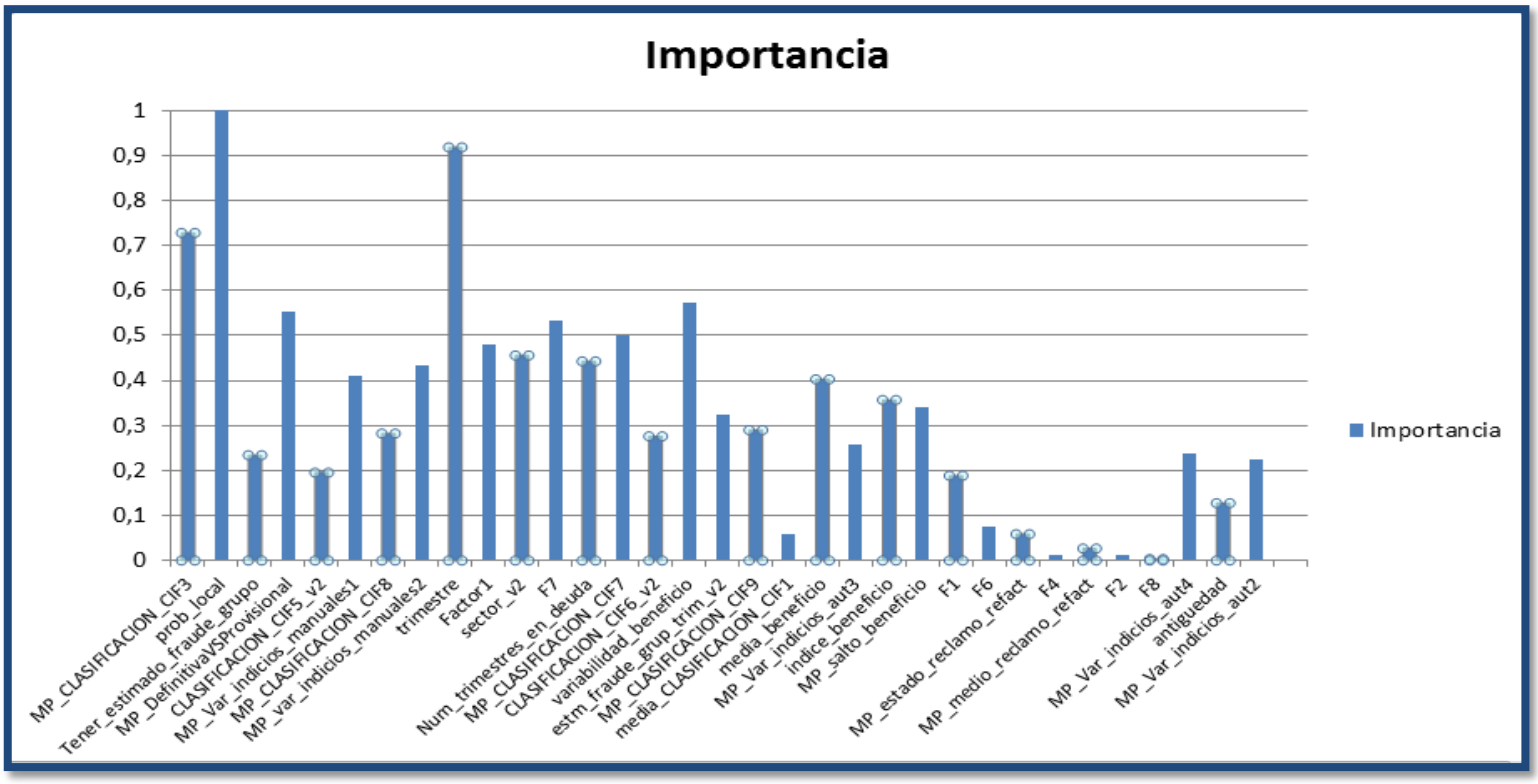


Tabla 6.5.2.1 1Importancia de las variables en el algoritmo de R.F.

Podemos apreciar que la variable con mayor importancia en la red es 'Prob_local', la cual, determina la probabilidad de fraude asignada a cada municipio. En contra posición encontramos que la variable menos influyente en la red es 'F8', la cual indicaba tener una incidencia reclamada por el sector 8 de la Agencia Tributaria.

6.5.3. Conclusión.

Se ha llevado a cabo una exhaustiva búsqueda de la mejor estructura de Random Forest, en el que se probaron un total de 11070 modelos, resultado de combinar un conjunto de parámetros con diferente parrilla de valores. La estimación del modelo se llevó a cabo a partir de los datos de Entrenamiento o Aprendizaje, mientras que la comparación o la evaluación de dichos modelos se realizaron sobre el conjunto de datos de Validación.

Por último y para concluir con este este apartado se va a proporcionar una estimación insesgada del grado de acierto de nuestro modelo óptimo, para lo cual, se utilizará el conjunto de datos Test.

Se muestra a continuación una tabla de contingencia entre los valores observados y predichos por el modelo para la variable dependiente Fraude:

		Fraude Estimado	
		-1	1
Fraude	-1	28484	5506
	1	14091	16001

- La probabilidad de clasificar de forma correcta a una empresa cualquiera registrada ante la Agencia Tributaria es del 0,69.
- La probabilidad de clasificar de forma correcta a una empresa no fraudulenta registrada ante la Agencia Tributaria es del 0,83 (especificidad).
- La probabilidad de clasificar de forma correcta a una empresa fraudulenta registrada ante la Agencia Tributaria es del 0,53 (sensibilidad).

6.6. Gradient Boosting.

6.6.1. Construcción del Gradient Boosting.

La construcción del mejor Gradient Boosting se lleva a cabo de forma empírica en un proceso de prueba-error donde se crea una parrilla de valores para los diferentes parámetros propios de esta técnica. Calculándose el coeficiente de correlación Matthews para cada una de las Redes construidas.

La parrilla de valores para los diferentes parámetros es:

- P-valor necesario para generar una regla de división (0.01, 0.05 y 0.1).
- Tamaño de hoja mínimo (100, 600, 1100, 1600 y 2100).
- Máximo número de árboles a crear (10, 50 y 100).
- Parámetro de Regularización (0.01, 0.05, 0.09, 1.30 y 1.70).
- Iteraciones (10, 50 y 100)
- Punto de corte. El modelo de Gradient Boosting clasificará como usuario fraudulento a aquellos que tengan una probabilidad de cometer fraude superior al punto de corte, este valor no es un parámetro inherente a la construcción de nuestras estructuras de Gradient Boosting, sino que se han calculado a partir de los datos de validación para evaluar el comportamiento o la efectividad de los modelos. El punto de corte tomará todos los valores entre 0.5 y 0.91 equidistantes 0.01.

En definitiva se construyen un total de 27675 modelos de Gradient Boosting resultado de todas las posibles combinaciones de la parrilla de valores.

Se muestra a continuación la cabecera de una tabla resumen para comparar los diferentes modelos de Gradient Boosting creados:

	VN	FP	FN	VP	Numrules	Matthews	PuntoDeCorteMin	maxtrees	leafsize	shrinkage	iterations	p_valor
1	82351	22073	45716	70942	240403	0.4012076746	0.51	100	100	0.05	100	0.05
2	82351	22073	45716	70942	240403	0.4012076746	0.51	50	100	0.05	100	0.01
3	82351	22073	45716	70942	240403	0.4012076746	0.51	10	100	0.05	100	0.05
4	84187	20237	48110	68548	49534	0.4010421167	0.51	100	600	0.05	100	0.05
5	84187	20237	48110	68548	49534	0.4010421167	0.51	50	600	0.05	100	0.05
6	84187	20237	48110	68548	49534	0.4010421167	0.51	10	600	0.05	100	0.05
7	85094	19330	49319	67339	49534	0.4009847757	0.52	100	600	0.05	100	0.01
8	85094	19330	49319	67339	49534	0.4009847757	0.52	50	600	0.05	100	0.01
9	85094	19330	49319	67339	49534	0.4009847757	0.52	10	600	0.05	100	0.05
10	83232	21192	46911	69747	240403	0.4006788738	0.52	100	100	0.05	100	0.01

Tabla 6.5.1.1. Resumen Modelos Gradient Boosting.

Se obtiene que el valor máximo en el coeficiente de correlación de Matthews (0.4012) es conseguido por el modelo de Gradient Boosting con los siguientes parámetros:

- P-valor necesario para generar una regla de división: 0.05.
- Tamaño de hoja mínimo: 100.
- Máximo número de árboles a crear: 100
- Parámetro de Regularización: 0.05.
- Iteraciones: 100.
- Punto de corte: 0.51

Al igual que pasaba con los modelos de Gradient Boosting, los diferentes modelos creados aparentemente se muestran bastante estables ante las variaciones propuestas de los parámetros estructurales “Número máximo de árboles” y “P-valor”.

6.6.2. Interpretación del modelo.

Importancia de las variables independientes: La importancia de una variable independiente es una medida que indica cuánto cambia el valor pronosticado por el modelo de Gradient Boosting para diferentes valores de la variable independiente. La importancia normalizada es el resultado de los valores de importancia divididos por el valor de importancia mayor, expresados como porcentajes.

Se muestra a continuación un gráfico de barras de los valores de la tabla de importancia, clasificado en valor de importancia descendente. La tabla que recoge los valores de importancia e importancia normalizada se encuentra en el Anexo de Gradient Boosting.

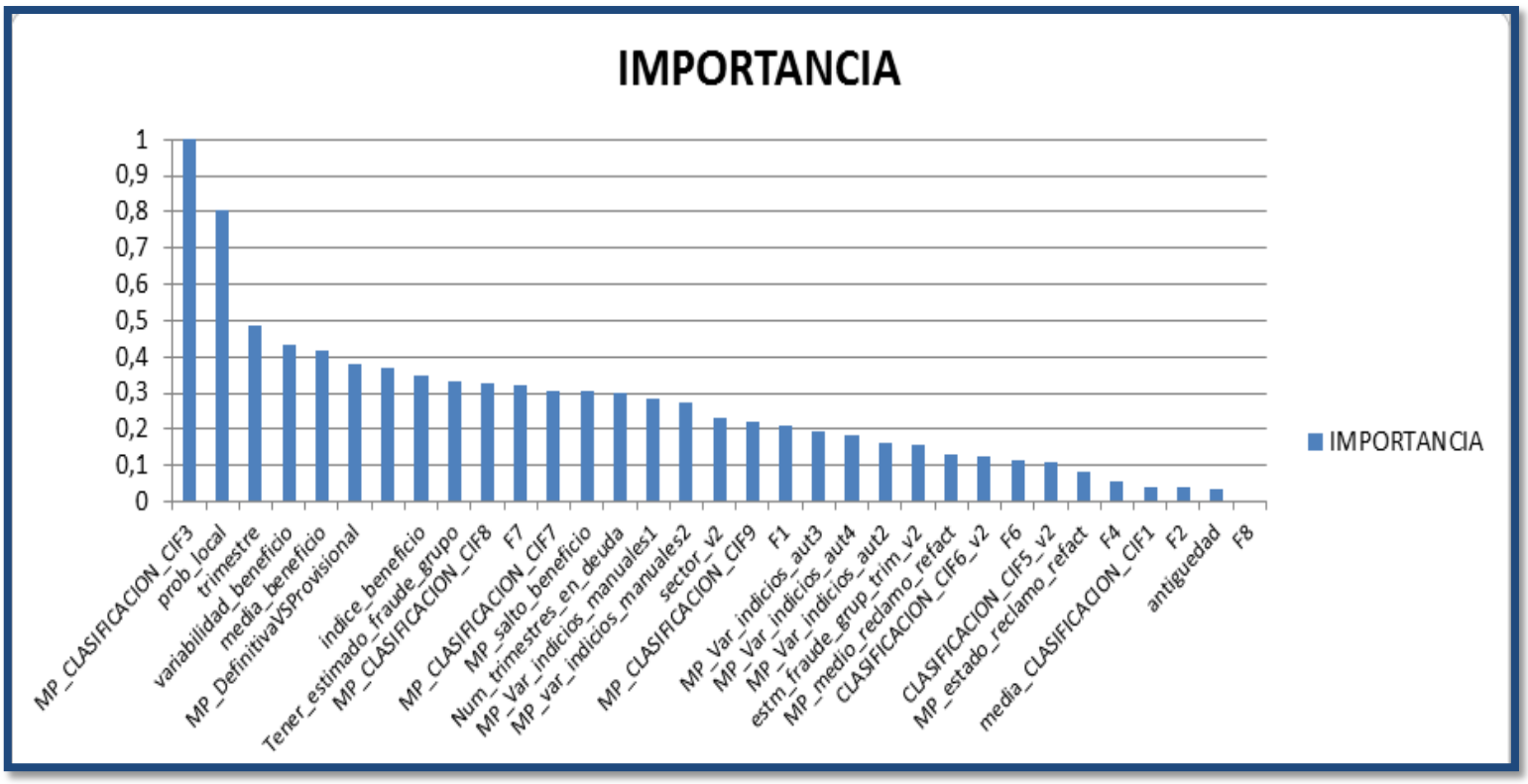


Tabla 6.6.2.1 21Importancia de las variables en el algoritmo de G.B.

Podemos apreciar que la variable con mayor importancia en la red es 'MP_Clasificacion_CIF3', la cual, determina la media ponderada de los 10 últimos trimestres de la tercera variable de clasificación del CIF de cada empresa. La media corresponde al fraude relativo de la variable respuesta para cada una de las categorías de la tercera variable de clasificación del CIF agrupada. En contra posición encontramos que la variable menos influyente en la red es 'F8', la cual indicaba tener una incidencia reclamada por el sector 8 de la Agencia Tributaria.

6.6.3. Conclusión.

Se ha llevado a cabo una exhaustiva búsqueda de la mejor estructura de Gradient Boosting, en el que se probaron un total de 27675 modelos, resultado de combinar un conjunto de parámetros con diferente parrilla de valores. La estimación del modelo se llevó a cabo a partir de los datos de Entrenamiento o Aprendizaje, mientras que la comparación o la evaluación de dichos modelos se realizaron sobre el conjunto de datos de Validación.

Por último y para concluir con este este apartado se va a proporcionar una estimación insesgada del grado de acierto de nuestro modelo óptimo, para lo cual, se utilizará el conjunto de datos Test. Se muestra a continuación una tabla de contingencia entre los valores observados y predichos por el modelo para la variable dependiente Fraude:

		Fraude Estimado	
		-1	1
Fraude	-1	26147	4747
	1	16028	17160

- La probabilidad de clasificar de forma correcta a una empresa cualquiera registrada ante la Agencia Tributaria es del 0,67.
- La probabilidad de clasificar de forma correcta a una empresa no fraudulenta registrada ante la Agencia Tributaria es del 0,84 (especificidad).
- La probabilidad de clasificar de forma correcta a una empresa fraudulenta registrada ante la Agencia Tributaria es del 0,51 (sensibilidad).

7. Ensamble de Modelos.

Como último apartado del estudio de modelos, se proponen aquí algunos métodos de ensamble de los modelos anteriormente ajustados con el fin de obtener clasificadores mejores (en términos del coeficiente de correlación Matthews) que los proporcionados por los modelos individuales.

Para ello se construye un conjunto de datos que contiene las probabilidades estimadas por los cuatro modelos ajustados con el fin de combinar estas probabilidades de distintas formas. Se construyen primero un total de 264 modelos de ensamble, resultado de combinar por un lado todos los posibles modelos creados (combinaciones de orden dos, tres y cuatro), combinándolos a su vez con las cuatro formas de agregar la información de las predicciones expuestas en la metodología (mínimo, máximo, mediana y media). En cada una de estas combinaciones (44) el punto de corte utilizado para clasificar las probabilidades estimadas se varía desde 0.4 a 1 en intervalos equidistantes de 0.1, por lo que en definitiva se prueba con un total de 264 modelos de ensamble.

Se muestra a continuación la cabecera de una tabla resumen para comparar los diferentes modelos de Ensamble creados:

	VP	FP	FN	VN	Matthews	Logistica	Redes	Forest	Boosting	Agregacion	p_corte
1	83216	21208	45162	71496	0.4145782889	0	1	0	1	median	0.5
2	83216	21208	45162	71496	0.4145782889	0	1	0	1	mean	0.5
3	84089	20335	46805	69853	0.4104476845	0	1	0	1	max	0.6
4	83913	20511	46677	69981	0.4096371166	1	1	0	1	max	0.6
5	74353	30071	35315	81343	0.408692893	0	1	0	1	max	0.5
6	84579	19845	47932	68726	0.4065809291	1	1	0	1	mean	0.5
7	71833	32591	32823	83835	0.4064900633	0	1	0	1	median	0.4
8	71833	32591	32823	83835	0.4064900633	0	1	0	1	mean	0.4
9	73741	30683	35001	81657	0.4055696233	1	1	0	1	max	0.5
10	73618	30806	35211	81447	0.4025907355	1	1	0	1	mean	0.4

Tabla 6.7.1. Resumen Modelos de Ensamble.

Se obtiene que el valor máximo en el coeficiente de correlación de Matthews (0.4145) es conseguido por el modelo de Ensamble formado por la combinación de los modelos de Redes Neuronales y Gradient Boosting utilizando la media o la mediana (mismos resultados) como forma de agregar la información de ambos modelos.

Se procede a calcular ahora otro conjunto de modelos de ensamble, los cuales se crean a partir de la Regresión Logística utilizando las probabilidades estimadas por cada una de las cuatro técnicas de predicción utilizadas. Como ya se describió en la metodología (apartado 2.7.) se calculan dos nuevas predicciones de probabilidad de fraude, una de ellas son directamente las probabilidades estimadas en la Regresión Logística utilizando como variables independientes las estimaciones de los modelos hasta ahora utilizadas (**Ensamble_logistica1**). Las otras nuevas predicciones se calculan realizando una media ponderada de las probabilidades estimadas. Dicha ponderación consiste en evaluar la importancia de cada predicción en función del valor relativo de su coeficiente de Regresión (**Ensamble_logistica2**). El punto de corte utilizado para clasificar las probabilidades estimadas se vuelve a variar desde 0.4 a 1 en intervalos equidistantes de 0.1, por lo que en definitiva se prueba con un total de 12 nuevos modelos de ensamble.

Se muestra a continuación la cabecera de una tabla resumen para comparar los diferentes modelos de Ensamble creados:

VP	FP	FN	VN	Matthews	P_corte	Ensamble_logistica
85247	19177	48547	68111	0,408738821	0,5	2
71833	32591	33390	83268	0,40152561	0,4	2
93878	10546	64357	52301	0,38442624	0,6	2
99759	4665	82650	34008	0,32438068	0,7	2
74223	30201	60182	56476	0,199301923	0,6	1
79942	24482	67384	49274	0,19898467	0,7	1
68139	36285	53493	63165	0,194661991	0,5	1
85689	18735	75758	40900	0,192555867	0,8	1
103562	862	105686	10972	0,190299485	0,8	2
61518	42906	46657	70001	0,188924012	0,4	1
92334	12090	87168	29490	0,175041029	0,9	1
104420	4	116401	257	0,031471526	0,9	2

Tabla 6.7.2. Resumen Modelos de Ensamble (Logística).

Se obtiene que el valor máximo en el coeficiente de correlación de Matthews (0.4087) es conseguido por el modelo de Ensamble construido mediante la segunda forma descrita en este apartado (media ponderada) y un punto de corte de 0.5. Cabe destacar que los modelos de Ensable_logistica1 no son "buenos", donde el mayor valor de Matthews no alcanza ni un valor de 0.20.

En definitiva los valores alcanzados mediante esta segunda tanda de modelos de ensamble no consiguen alcanzar los resultados conseguidos por el modelo de ensamble formado por Redes Neuronales y Gradient Boosting.

Para concluir con este apartado se va a proporcionar una estimación insesgada del grado de acierto de nuestro modelo de ensamble óptimo, para lo cual, se utilizará el conjunto de datos Test. Se muestra a continuación una tabla de contingencia entre los valores observados y predichos por el modelo para la variable dependiente Fraude:

		Fraude Estimado	
		-1	1
Fraude	-1	27691	5130
	1	14484	16777

- La probabilidad de clasificar de forma correcta a una empresa cualquiera registrada ante la Agencia Tributaria es del 0,69.
- La probabilidad de clasificar de forma correcta a una empresa no fraudulenta registrada ante la Agencia Tributaria es del 0,84 (especificidad).
- La probabilidad de clasificar de forma correcta a una empresa fraudulenta registrada ante la Agencia Tributaria es del 0,53 (sensibilidad).

8. Comparación de modelos.

8.1. Introducción.

Una vez llegados a este punto queda determinar cuál ha sido el mejor modelo de predicción, cuál es la técnica estadística que mejor predice la clasificación de las empresas registradas ante la Agencia Tributaria.

Para tomar la decisión anterior se ha utilizado el conjunto de datos '*Validación*', donde sobre dicho conjunto de datos se han calculado diferentes resultados sobre el poder predictivo de cada uno de los modelos.

8.2. ¿Cuál es el mejor modelo de predicción?

Se muestra a continuación una tabla que recoge el valor del coeficiente de correlación de Matthews para cada uno de los modelos de clasificación óptimos de cada técnica utilizada, así como los valores necesarios para construir dicho coeficiente:

Modelo	VP	FP	FN	VN	Tasa de Aciertos	Matthews
Regresión Logística	76988	27436	41220	75438	0,689454592	0,3842
Redes Neuronales	79624	24800	44328	72330	0,687319637	0,3921
Random Forest	88661	15763	56172	60486	0,674622991	0,3860
Gradient Boosting	82351	22073	45716	70942	0,693376213	0,4012
Ensamble	81994	22430	44066	72592	0,699224722	0,4145

Tabla 7.2.1. Resumen Comparativo de Modelos.

Como podemos observar en la tabla anterior el máximo valor del Coeficiente de Correlación de Matthews (0.4145) se consigue mediante el modelo de ensamblado. Dicho modelo tiene una diferencia significativa con respecto al resto de modelos y más especialmente si lo comparamos con el modelo de Regresión Logística, el cual, es la única alternativa que ofrece la posibilidad de ser interpretado. Cabe destacar que con cualquiera de las técnicas utilizadas en el presente estudio se consigue una tasa de aciertos en la clasificación de un 70% aproximado.

En definitiva consideraremos el modelo de ensamble de Redes Neuronales y Gradient Boosting como el modelo óptimo para predecir la probabilidad de que una empresa este cometiendo fraude ante la Agencia Tributaria.

9. Post-Análisis.

Con el objetivo de poder entender en la medida de lo posible el comportamiento de nuestro modelo probabilístico, se procede a realizar un conjunto de histogramas que representen para cada una de las variables explicativas dos gráficos, el primero de ellos filtrando por aquellos usuarios a los que el modelo predice visitar y el segundo filtrando por aquellos a los que el modelo clasifica como no fraudulentos.

En muchos de estos Histogramas no se ha podido llevar a cabo una interpretación clara, la cual, pueda ser expuesta en el presente estudio. Esto se debe al carácter confidencial de los datos. No obstante se detallan a continuación algunos de los resultados más relevantes encontrados en este apartado. Los histogramas se encuentran representados en el Anexo 11.5.

El ensamble de modelos formado por Redes Neuronales y Gradient Boosting, el cual, se ha considerado la mejor técnica estadística para clasificar a las empresas registradas ante la Agencia tributaria se comporta siguiendo los siguientes patrones:

- Se tenderá a visitar más a aquellas empresas que pertenezcan a una zona especial (Clasificación_CIF6_V2=1).
- Se tenderá a visitar más a aquellas empresas que no tengan un fraude estimado para el grupo al que pertenezcan (Tener_estimado_fraude_grupo=0).
- Se tenderá a visitar más a aquellas empresas que sean consideradas a priori más fraudulentas que el resto (Clasificación_CIF5_V2=1).
- El ensamblado de modelos tratara por igual a los individuos que tengan valores bajos en las variables Factor1, MP_DefinitivaVSProvisional y Num_trimestres_en_deuda. No obstante si se tendera a visitar más a aquellos individuos que tengan valores medios y altos en dichas variables.
- En cuanto a la variable Prob_local, se produce una clara separación, donde se tendera a visitar más a aquellos usuarios con un valor inferior a 0.40.
- Para la variable Sector_V2 no se observan diferencias en términos de a quien si se decide visitar, no obstante los histogramas dejan intuir que aquellos con valores bajos en esta variable tenderán a ser considerados como no fraudulentos con más normalidad que aquellos que por el contrario no tomen valores bajos en dicha variable.
- En cuanto al resto de variables explicativas, los histogramas no muestran ninguna tendencia clara de comportamiento por parte del Ensamble de modelos.

10. Conclusiones.

El fraude empresarial es uno de los grandes problemas de nuestro País y no existe una fórmula secreta para detectar a las empresas que defraudan al estado. La estadística pese a no ser *'algo mágico'* tiene un gran poder predictivo en la gran mayoría de sucesos que se dan en la realidad. Como se ha visto en el presente estudio, con facilidad relativa se puede clasificar a una empresa con un porcentaje de éxito de entorno al 70%.

Para que las técnicas estadísticas ofrezcan buenos resultados es preciso por parte del investigador un amplio conocimiento de la información de partida. En el presente estudio dado la confidencialidad de los datos no se ha podido presentar de forma amplia y detalla la información con la que se ha partido, no obstante se ha requerido un amplio conocimiento de la misma para poder en el punto 5 llegar a la información tratada, la cual, ha sido la utilizada por el software estadísticos SAS para la construcción de los modelos de predicción. Sin dicho conocimiento de la información de partida, previo al comienzo de los análisis, los resultados esperados habrían sido mucho peores.

Antes de llevar a cabo ningún modelo de predicción hemos reducido la dimensionalidad de las siete variables Socio-Económicas en una sola variable que recogía casi el 90% de la información de las siete.

A la hora de realizar los modelos de predicción hemos logrado analizar los datos tanto en el conjunto de datos de aprendizaje para estimar los parámetros de cada modelo como en validación utilizado para lograr diferencia a un único modelo predictivo como el mejor de todos y en el conjunto de datos test para obtener una estimación insesgada de *"lo bueno"* que ha sido nuestro trabajo.

En cuanto a la regresión logística hemos conseguido diferenciar que variables eran significativas en el modelo y cuáles no, así como cuantificar como afecta a la probabilidad de cometer fraude el cambio en las variables regresoras (OR). Tratando de forma diferente aquellas que eran categóricas con respecto a las numéricas.

Por otro lado, en cuanto al resto de técnicas de predicción (al igual que se ha hecho con la Regresión Logística) la forma de encontrar la mejor estructura a sido de forma empírica, probando con todas las posibles combinaciones de una parrilla de valores. La construcción de dicha parrilla de valores se ha llevado a cabo intentando no dejar ninguna alternativa "fuera", aunque esto se ha tenido que ajustarse a las limitaciones impuestas por el hardware utilizado. En definitiva se han llevado a cabo un total de 41508 modelos de predicción.

En cuanto al modelo óptimo de predicción creado, este ha sido un ensamble de dos modelos, formado por un modelo de Redes Neuronales junto con un modelo de Gradient Boosting. Este modelo óptimo carece de la posibilidad de ser interpretado pero ha ofrecido un Coeficiente de Correlación de Matthews significativamente mayor que el resto.

Se procedió por ultimo a predecir de forma insesgada el grado de acierto de nuestro modelo óptimo, para ello se utilizará el conjunto de datos Test. Los resultados más relevantes fueron:

- La probabilidad de clasificar de forma correcta a una empresa cualquiera registrada ante la Agencia Tributaria es del 0,69.
- La probabilidad de clasificar de forma correcta a una empresa no fraudulenta registrada ante la Agencia Tributaria es del 0,84 (especificidad).
- La probabilidad de clasificar de forma correcta a una empresa fraudulenta registrada ante la Agencia Tributaria es del 0,53 (sensibilidad).

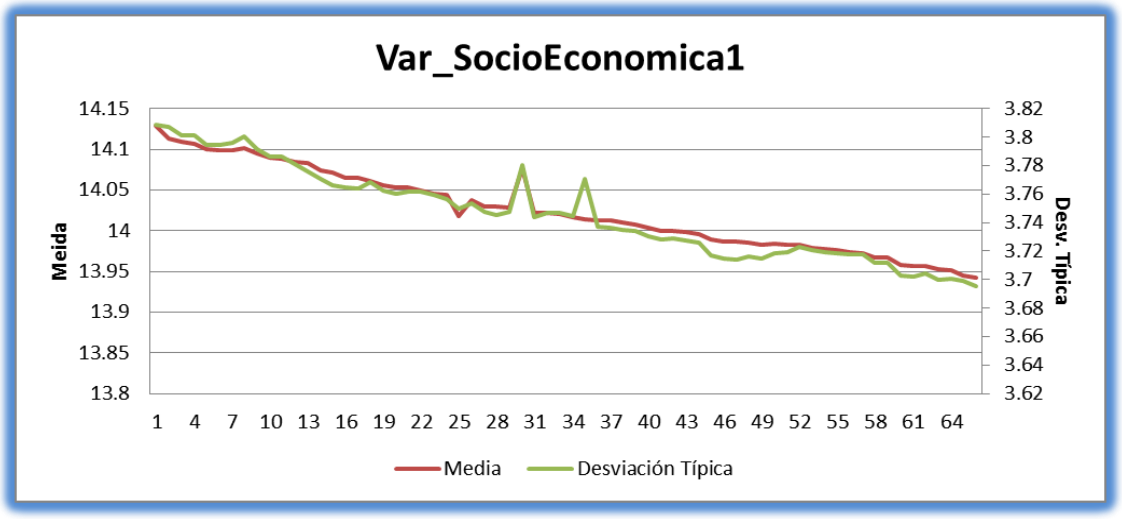
Estos resultados se han de considerar esperanzadores aunque mejorable en el tiempo ante la lucha contra el fraude empresarial.

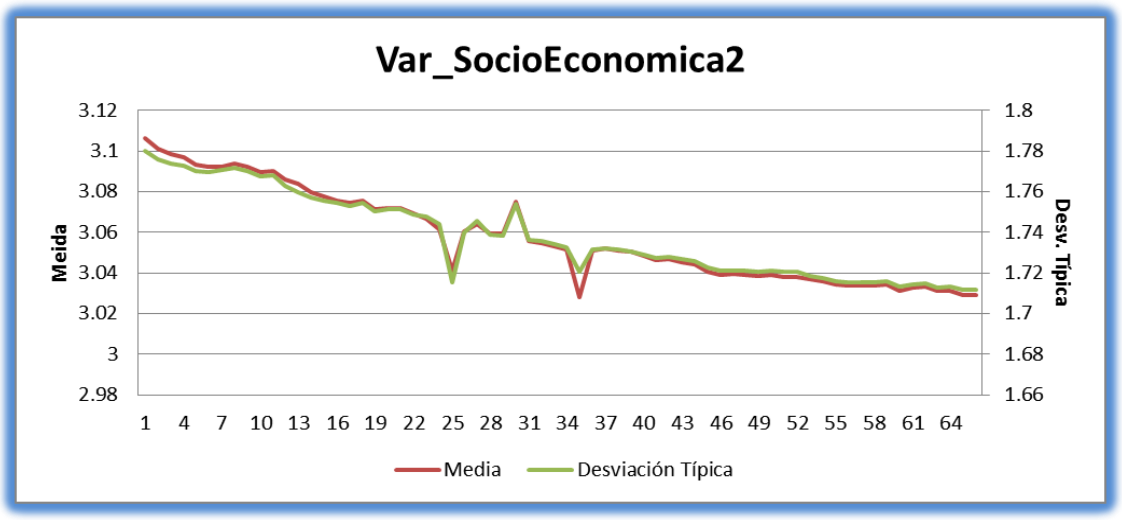
11. Anexos.

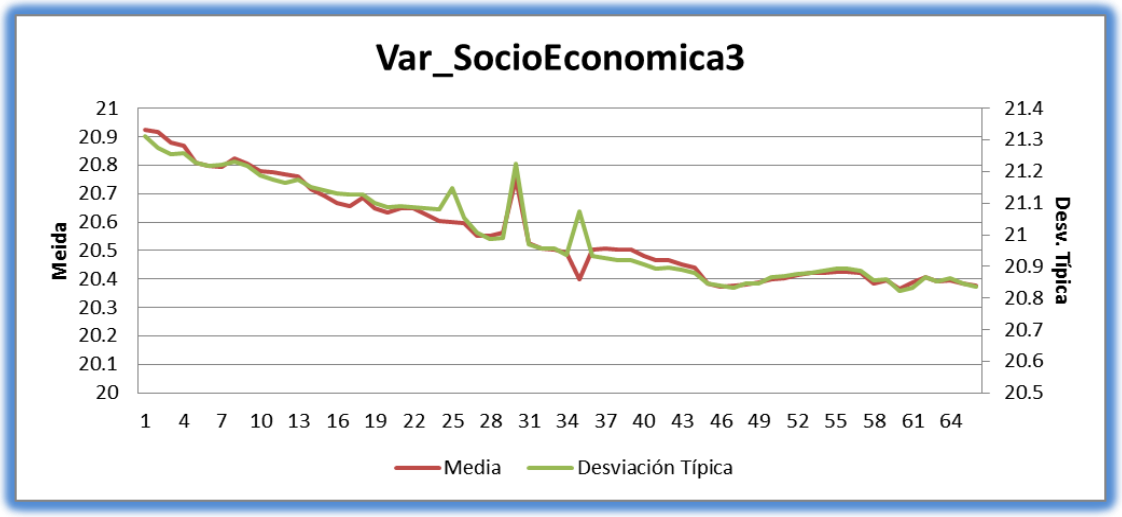
11.1. Anexo Descriptivos.

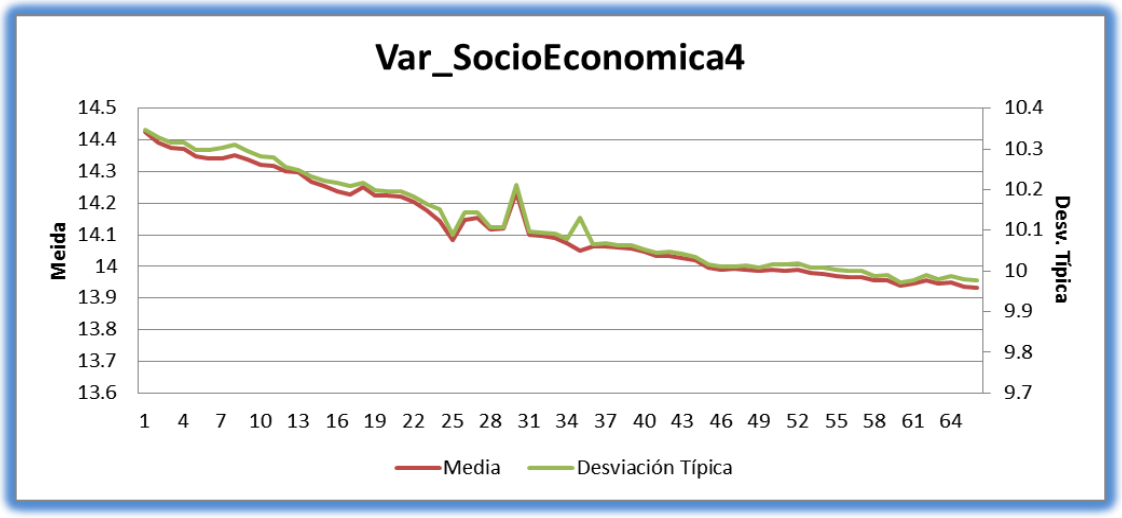
Los gráficos carecen de la explicación del porqué de dicho comportamiento no porque no se haya realizado sino porque desvelaría demasiada información del objeto real de estudio.

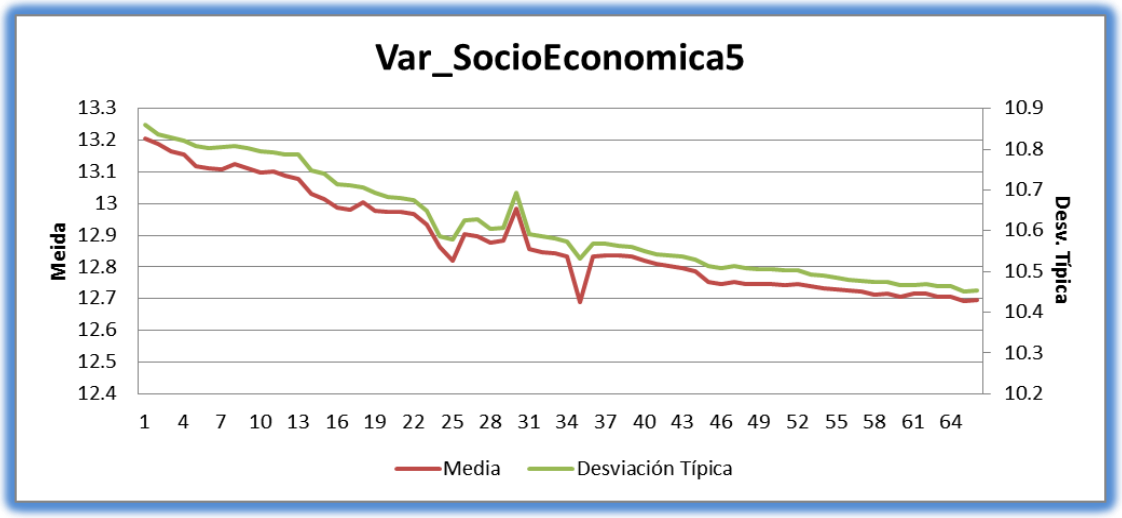
11.1.1. Variables Cuantitativas.

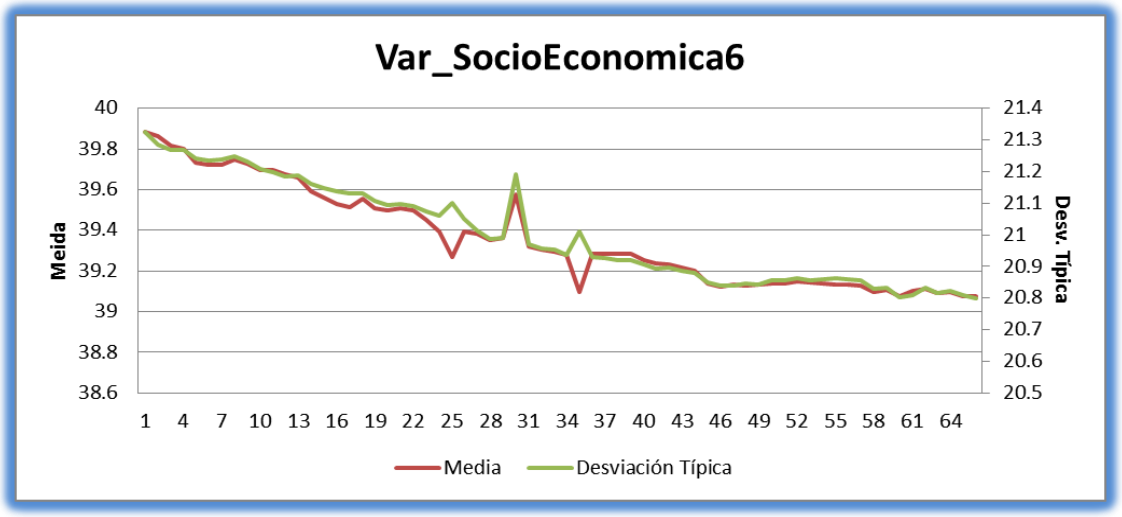


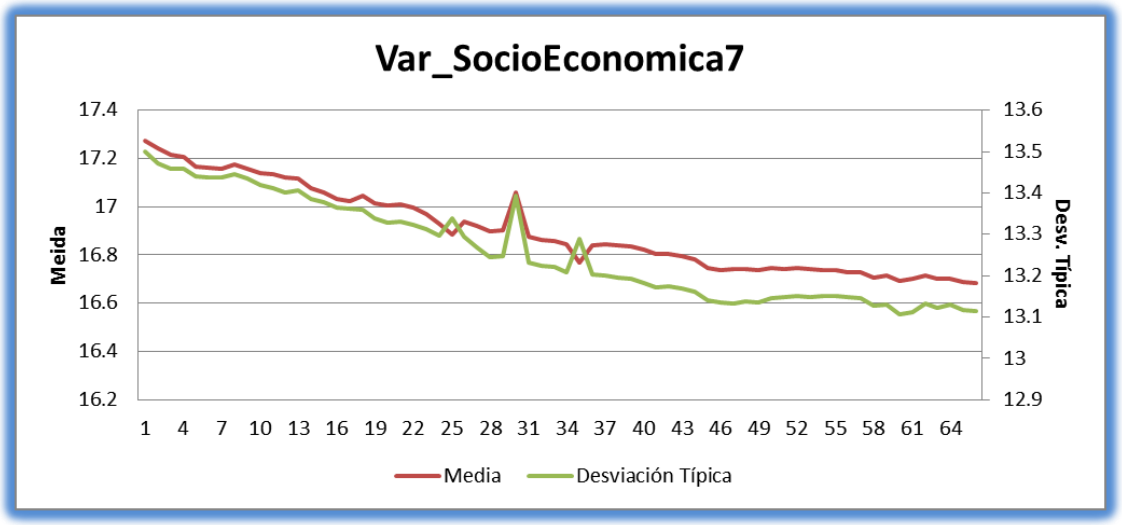


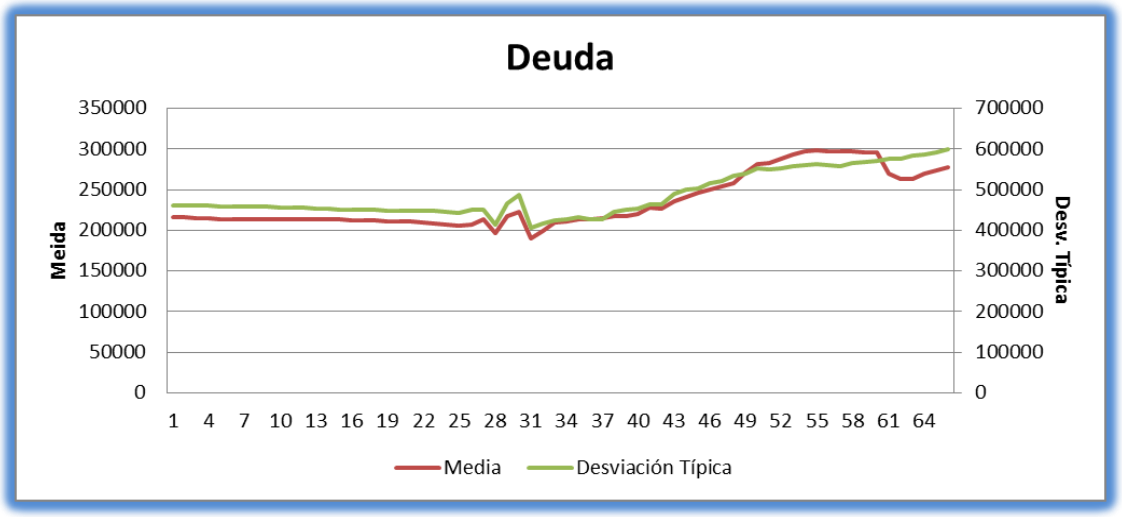

Var_SocioEconomica3
Meida
Desv. Típica
21
20.9
20.8
20.7
20.6
20.5
20.4
20.3
20.2
20.1
20
21.4
21.3
21.2
21.1
21
20.9
20.8
20.7
20.6
20.5
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

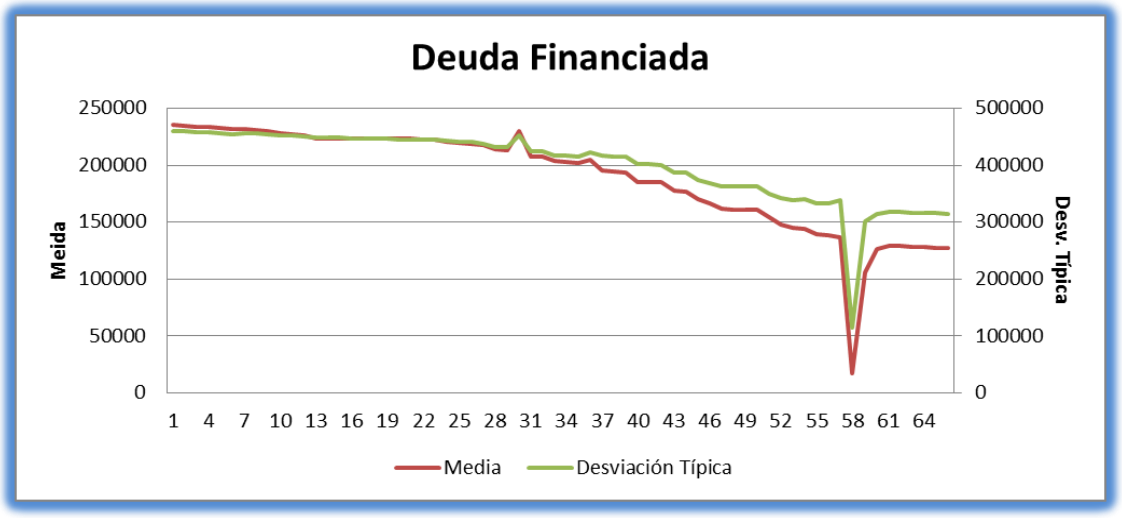

Var_SocioEconomica4
Meida
Desv. Típica
14.5
14.4
14.3
14.2
14.1
14
13.9
13.8
13.7
13.6
10.4
10.3
10.2
10.1
10
9.9
9.8
9.7
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

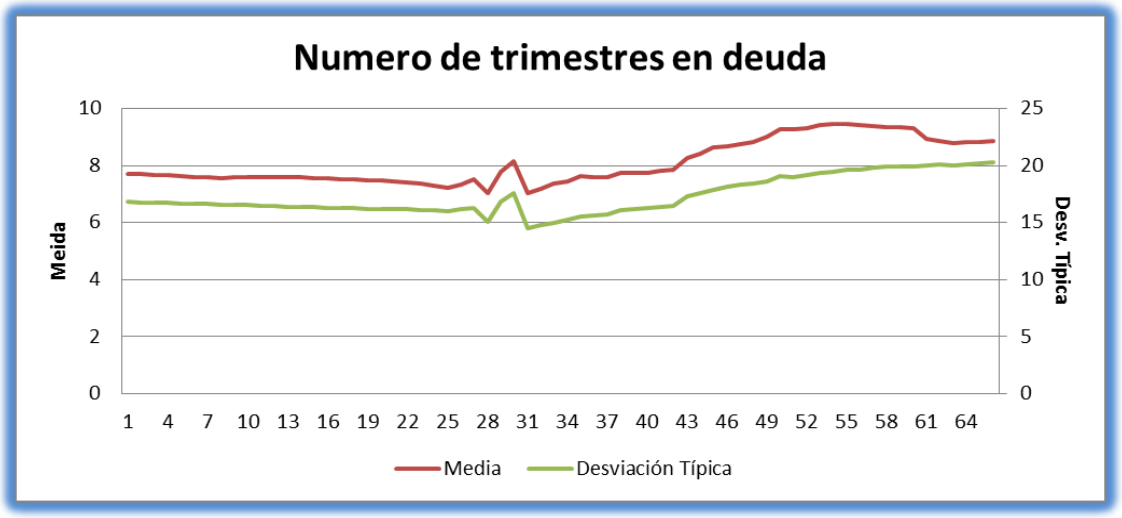

Var_SocioEconomica5
Meida
Desv. Típica
13.3
13.2
13.1
13
12.9
12.8
12.7
12.6
12.5
12.4
10.9
10.8
10.7
10.6
10.5
10.4
10.3
10.2
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

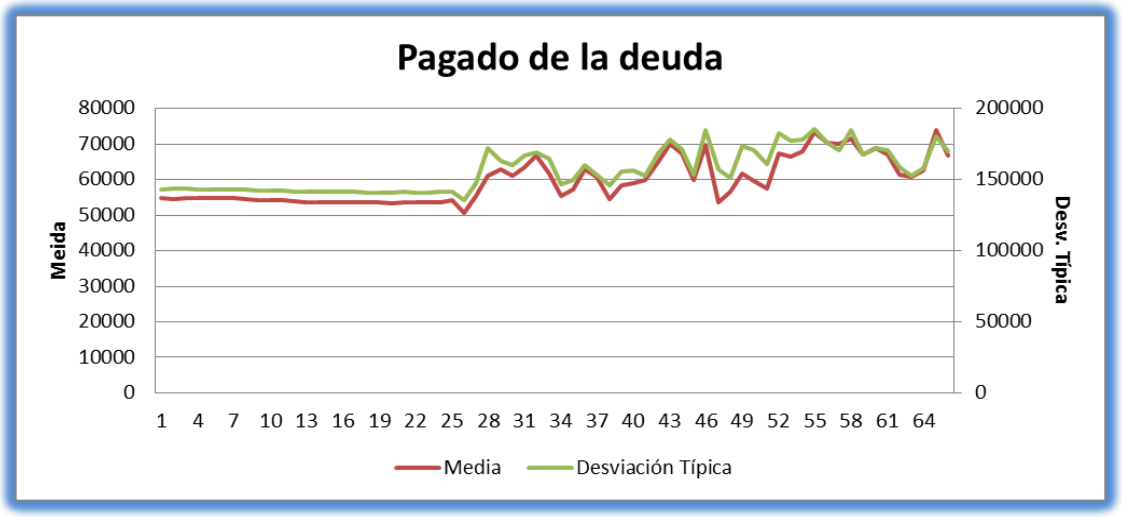

Var_SocioEconomica6
Meida
Desv. Típica
40
39.8
39.6
39.4
39.2
39
38.8
38.6
21.4
21.3
21.2
21.1
21
20.9
20.8
20.7
20.6
20.5
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

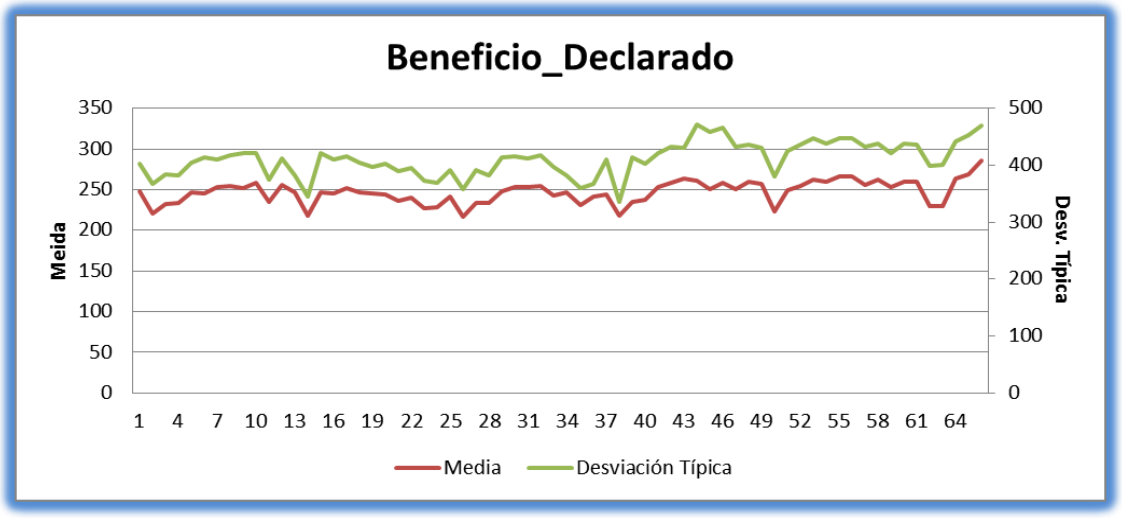

Var_SocioEconomica7
Meida
Desv. Típica
17.4
17.2
17
16.8
16.6
16.4
16.2
13.6
13.5
13.4
13.3
13.2
13.1
13
12.9
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

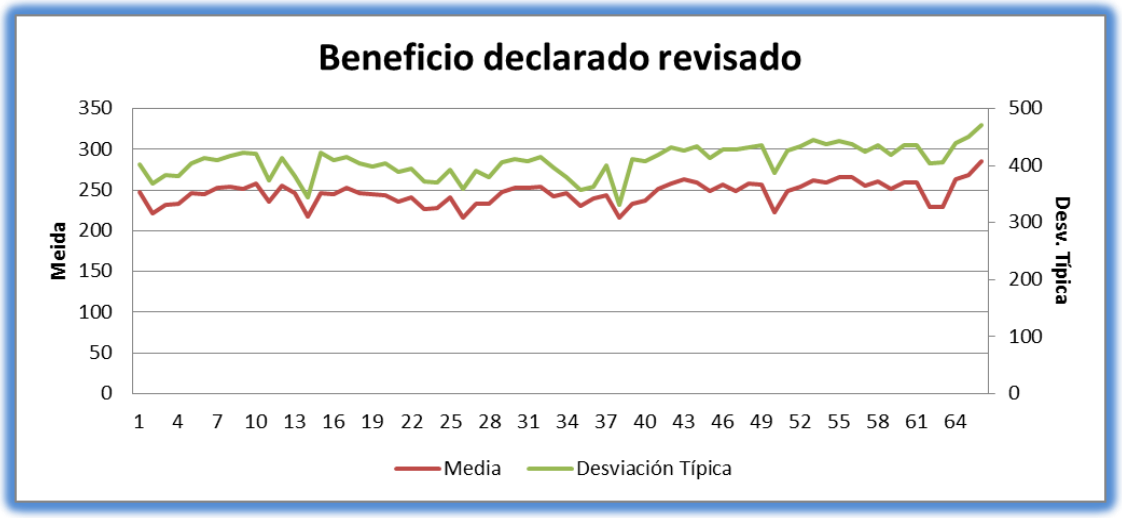

Deuda
Meida
Desv. Típica
350000
300000
250000
200000
150000
100000
50000
0
700000
600000
500000
400000
300000
200000
100000
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

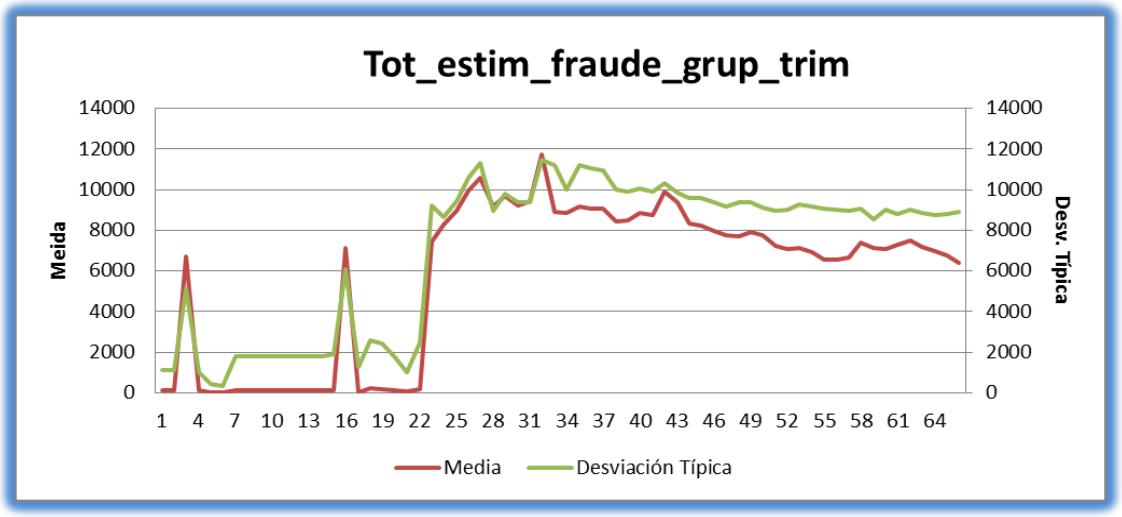

Deuda Financiada
Meida
Desv. Típica
250000
200000
150000
100000
50000
0
500000
400000
300000
200000
100000
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

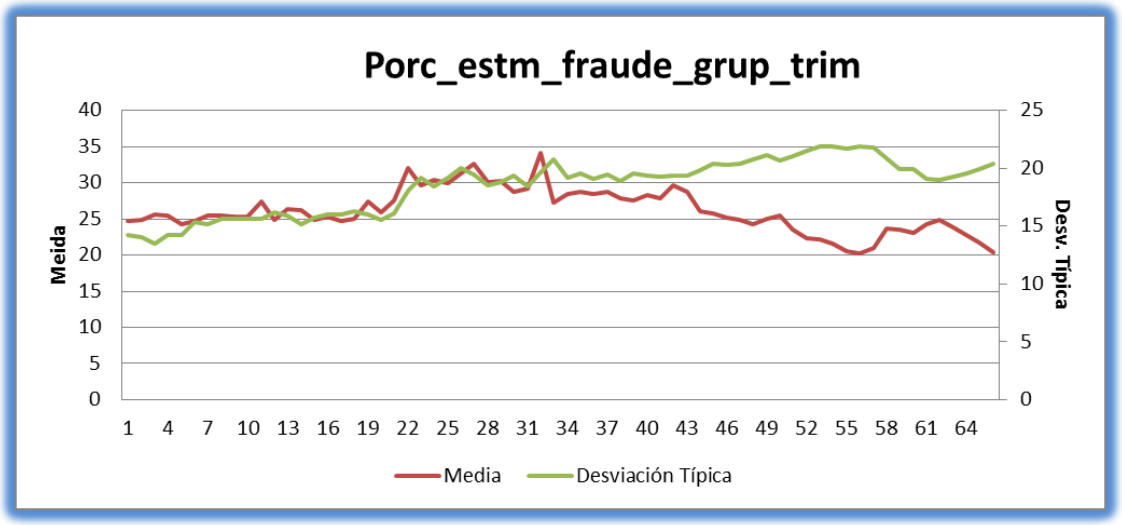

Numero de trimestres en deuda
Meida
Desv. Típica
10
8
6
4
2
0
25
20
15
10
5
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

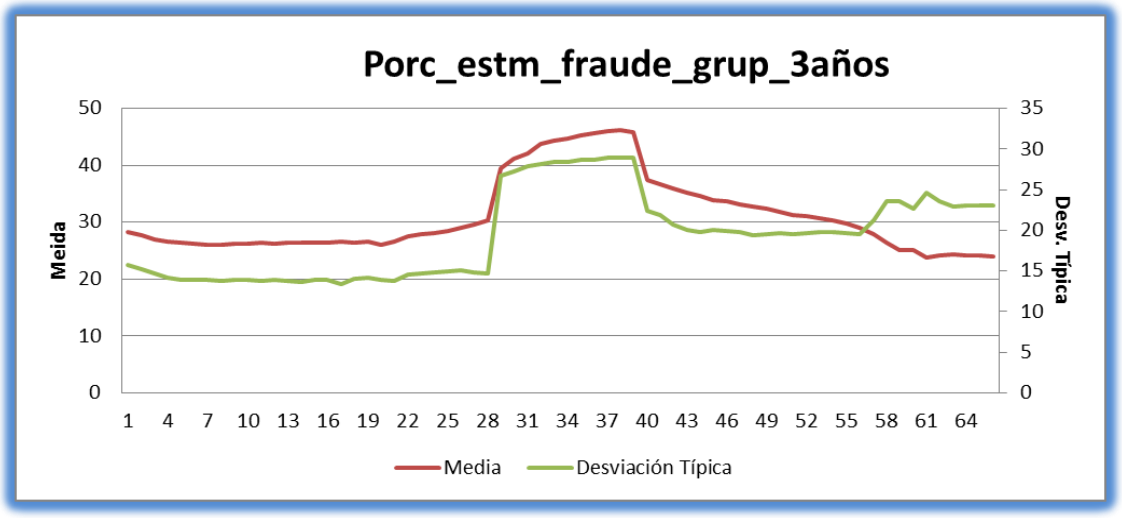

Pagado de la deuda
Meida
Desv. Típica
80000
70000
60000
50000
40000
30000
20000
10000
0
200000
150000
100000
50000
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica


Beneficio_Declarado
Meida
Desv. Típica
350
300
250
200
150
100
50
0
500
400
300
200
100
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

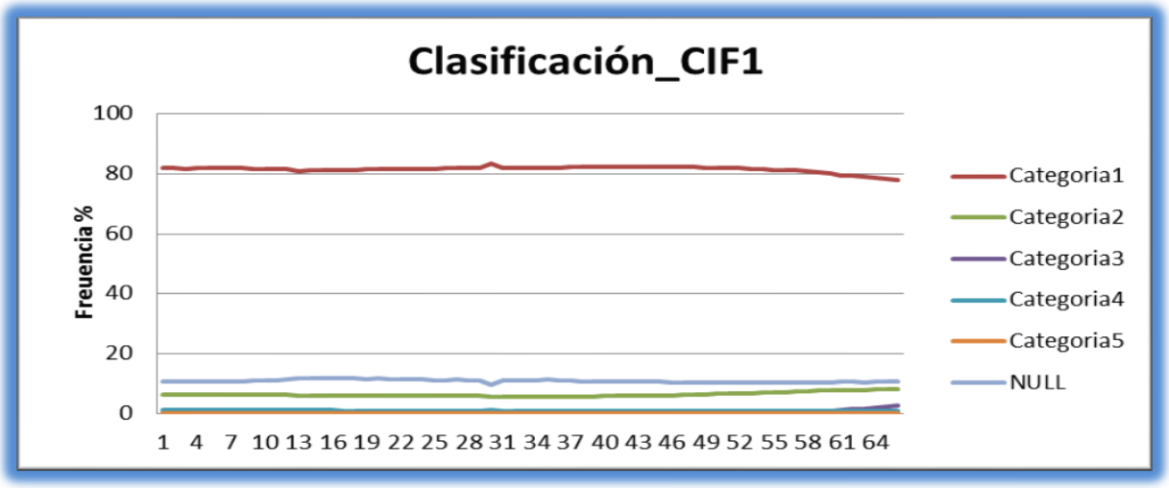

Beneficio declarado revisado
Meida
Desv. Típica
350
300
250
200
150
100
50
0
500
400
300
200
100
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

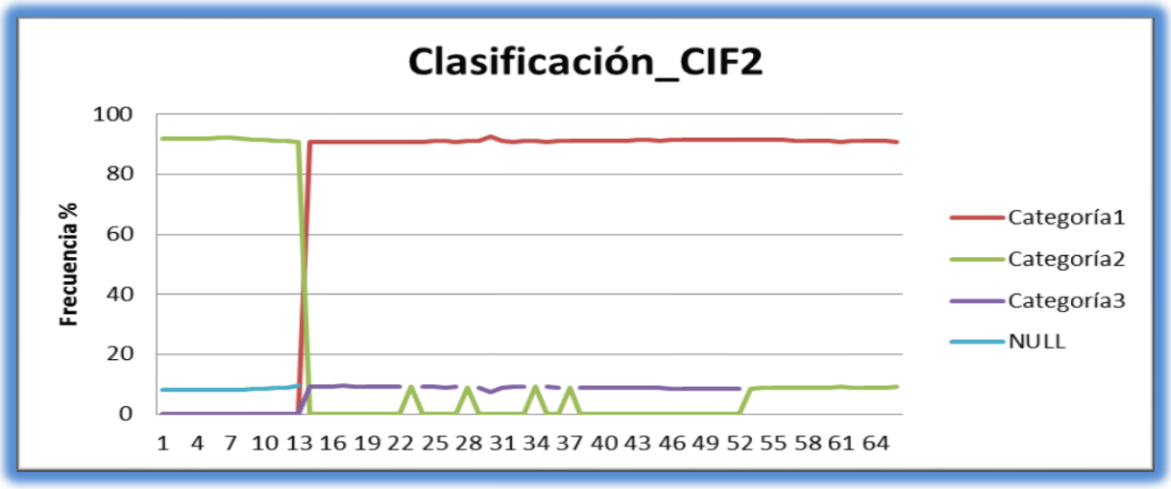

Tot_estim_fraude_grup_trim
Meida
Desv. Típica
14000
12000
10000
8000
6000
4000
2000
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

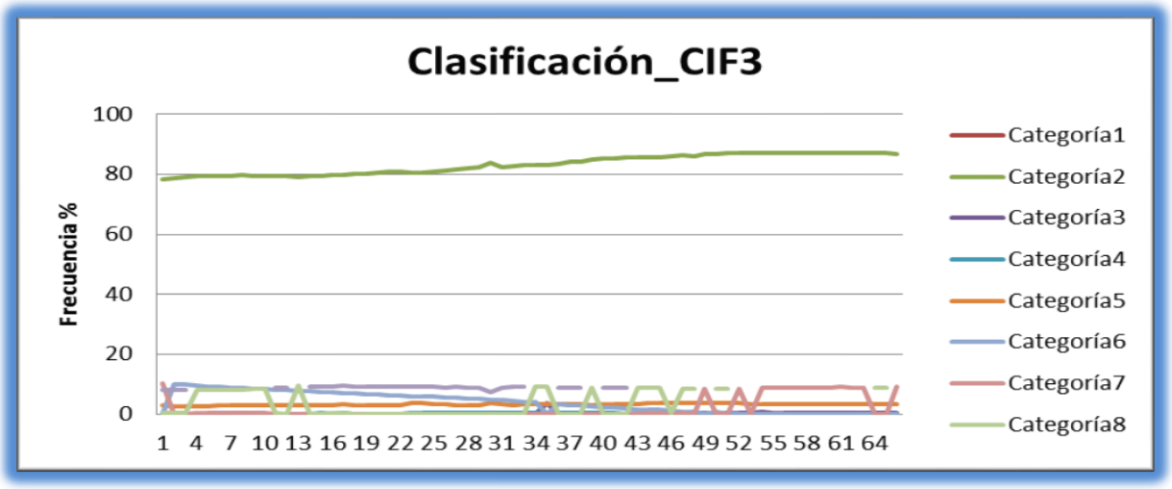

Porc_estm_fraude_grup_trim
Meida
Desv. Típica
40
35
30
25
20
15
10
5
0
25
20
15
10
5
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

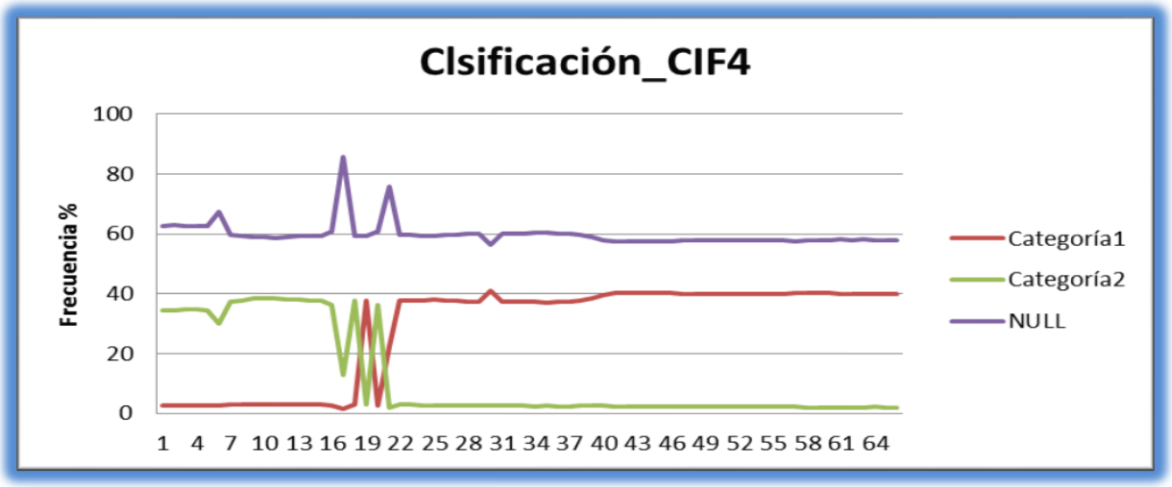

Porc_estm_fraude_grup_3años
Meida
Desv. Típica
50
40
30
20
10
0
35
30
25
20
15
10
5
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Media
Desviación Típica

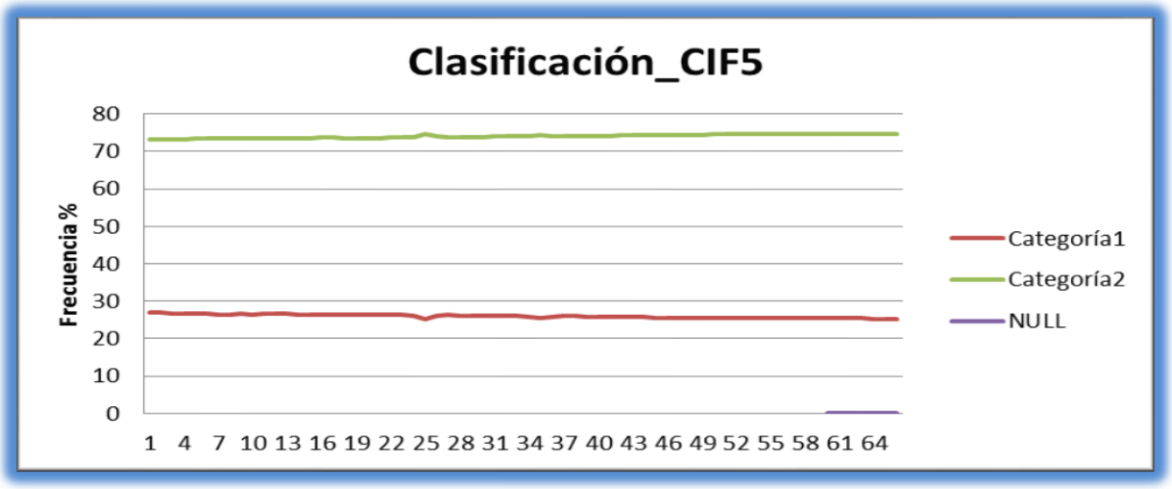
11.1.2. Variables Cualitativas.

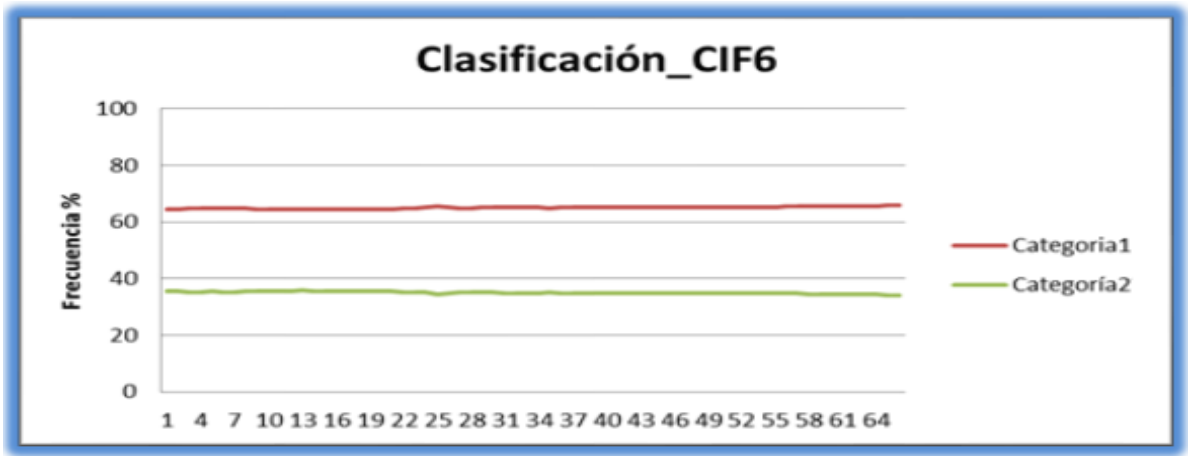


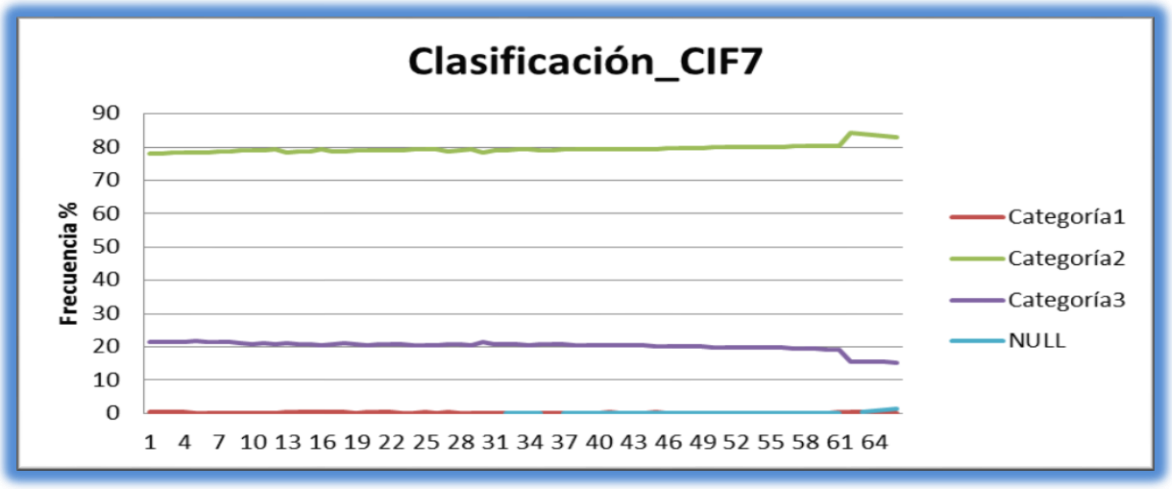


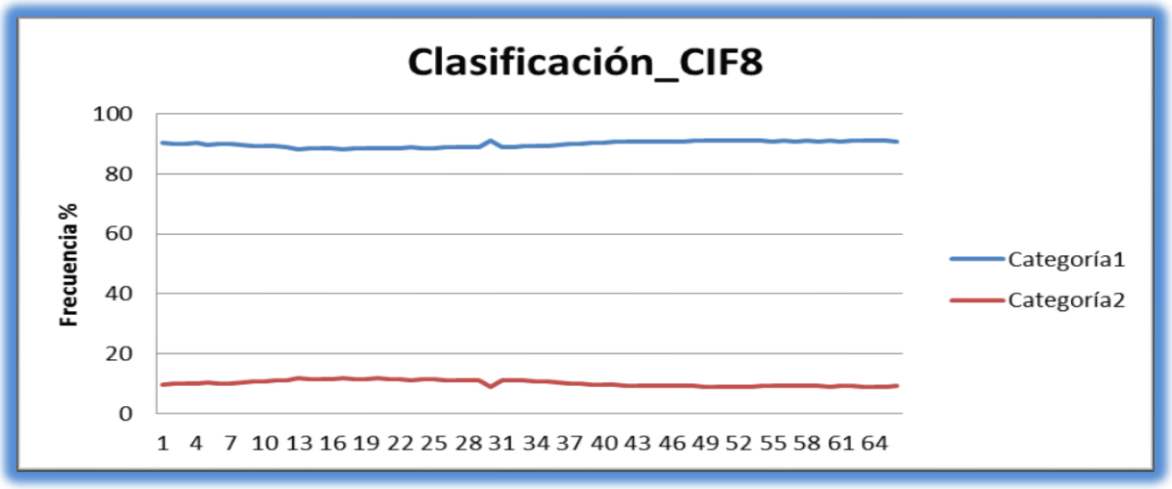

Clasificación_CIF3
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoría1
Categoría2
Categoría3
Categoría4
Categoría5
Categoría6
Categoría7
Categoría8

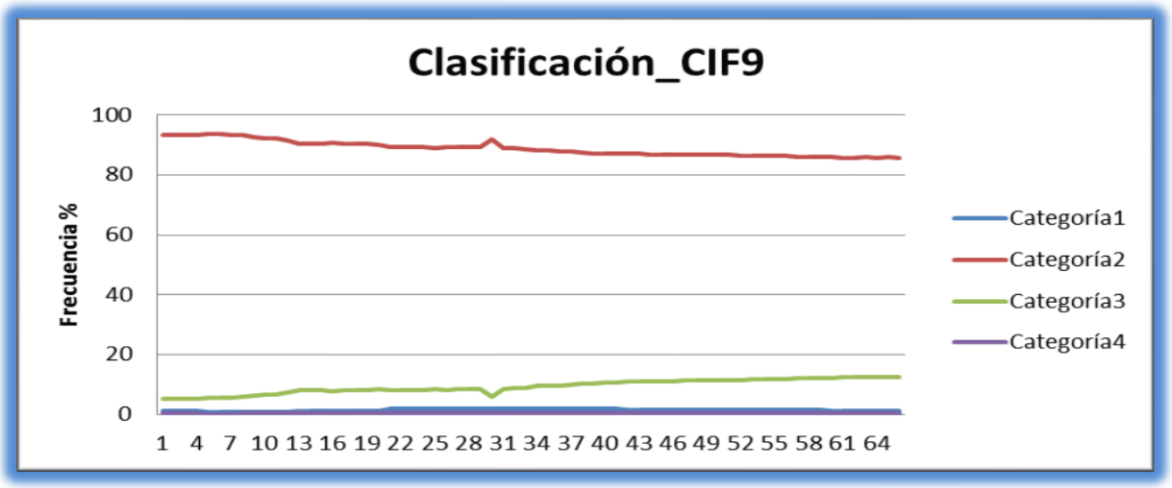

Clsificación_CIF4
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoría1
Categoría2
NULL

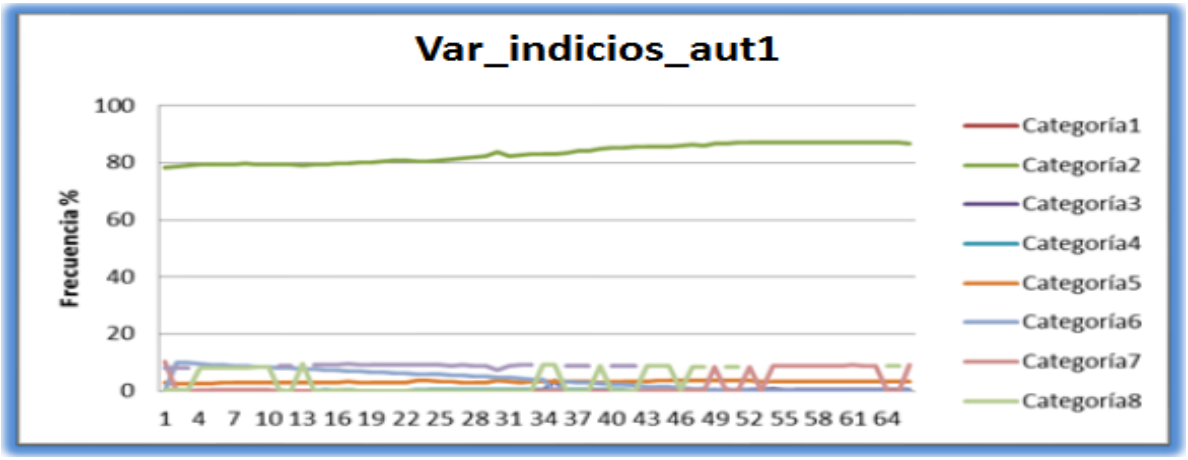

Clasificación_CIF5
Frecuencia %
80
70
60
50
40
30
20
10
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoría1
Categoría2
NULL

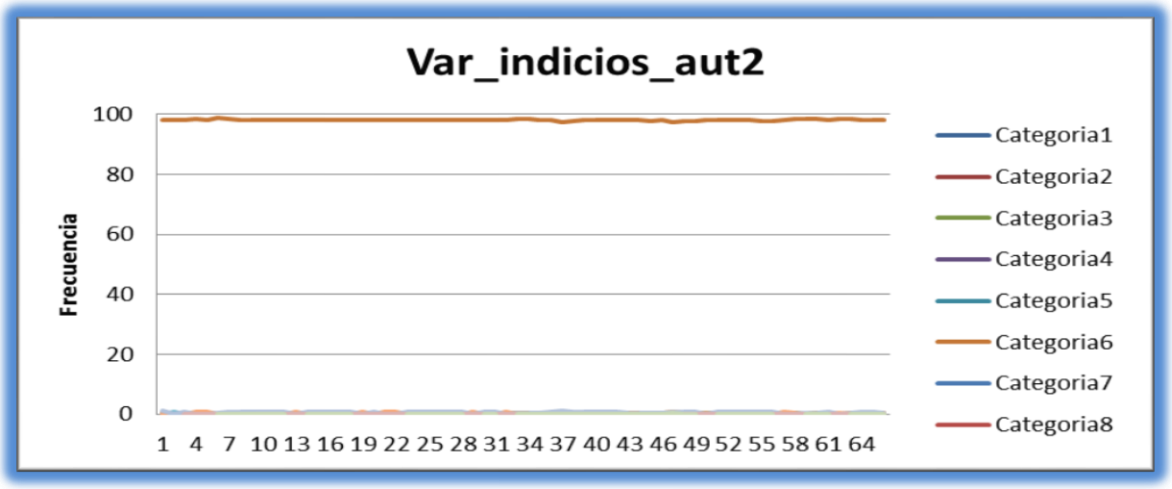

Clasificación_CIF6
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoria1
Categoría2

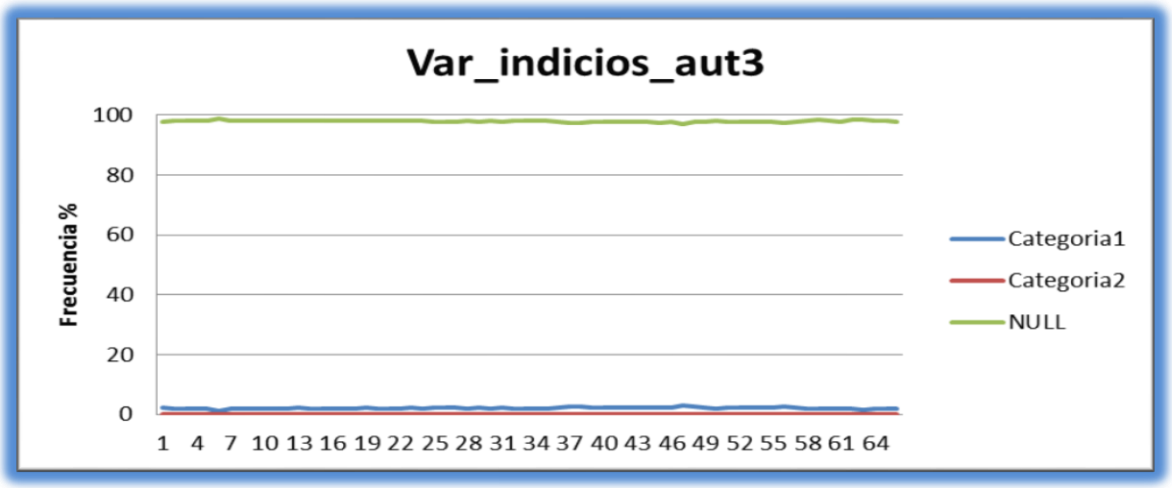

Clasificación_CIF7
Frecuencia %
90
80
70
60
50
40
30
20
10
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoría1
Categoría2
Categoría3
NULL

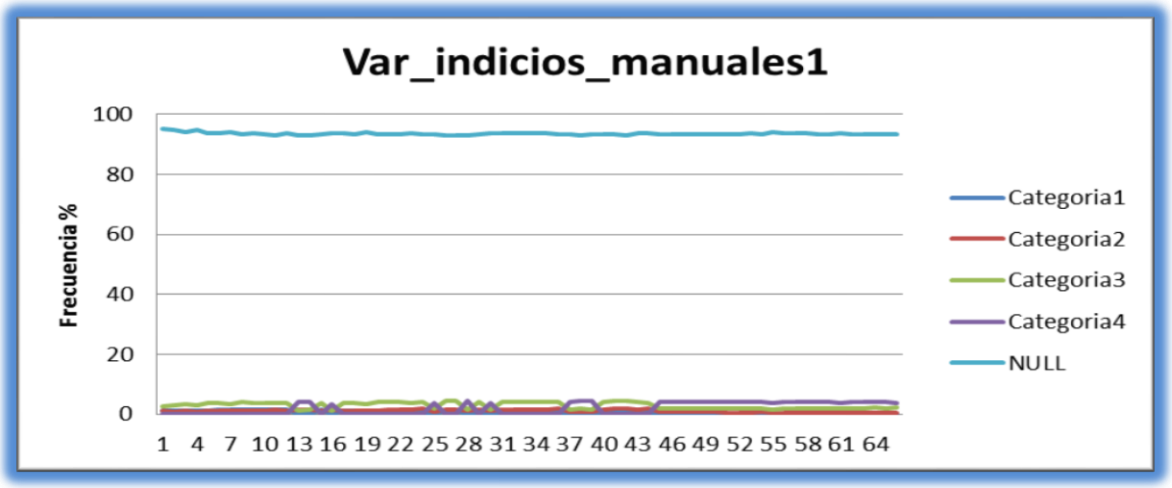

Clasificación_CIF8
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoría1
Categoría2

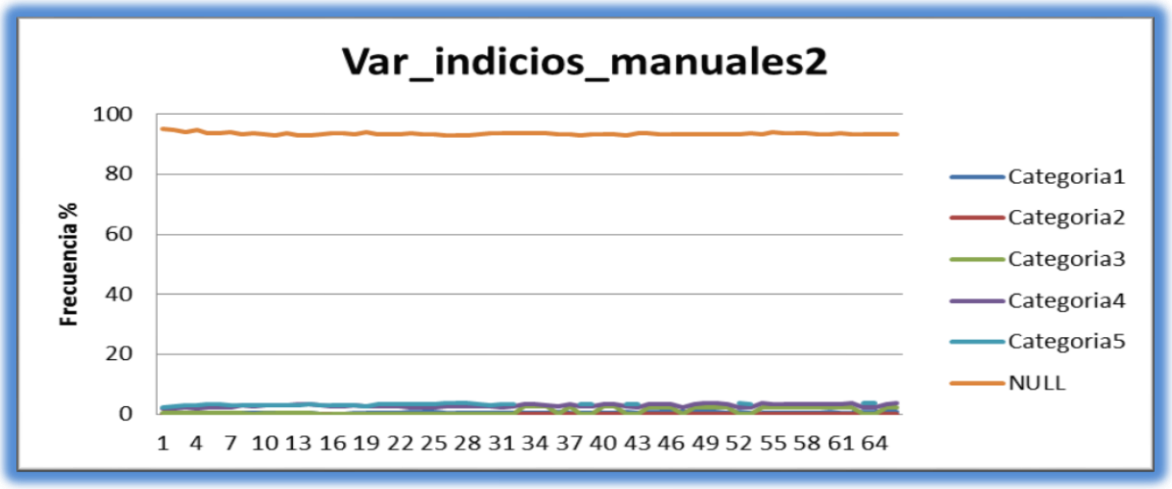

Clasificación_CIF9
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoría1
Categoría2
Categoría3
Categoría4

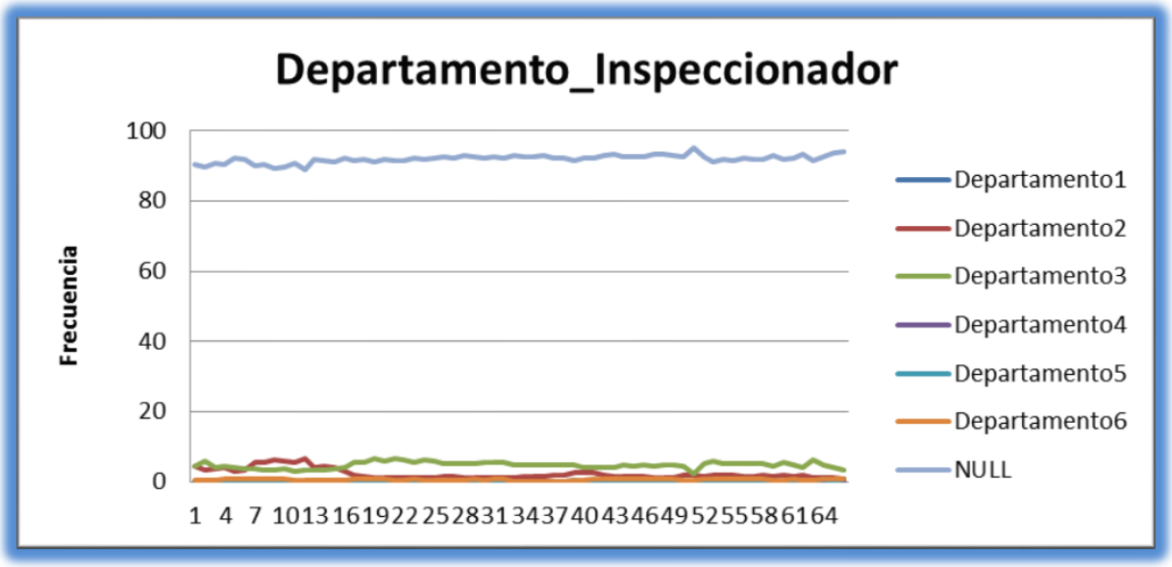

Var_indicios_aut1
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoría1
Categoría2
Categoría3
Categoría4
Categoría5
Categoría6
Categoría7
Categoría8

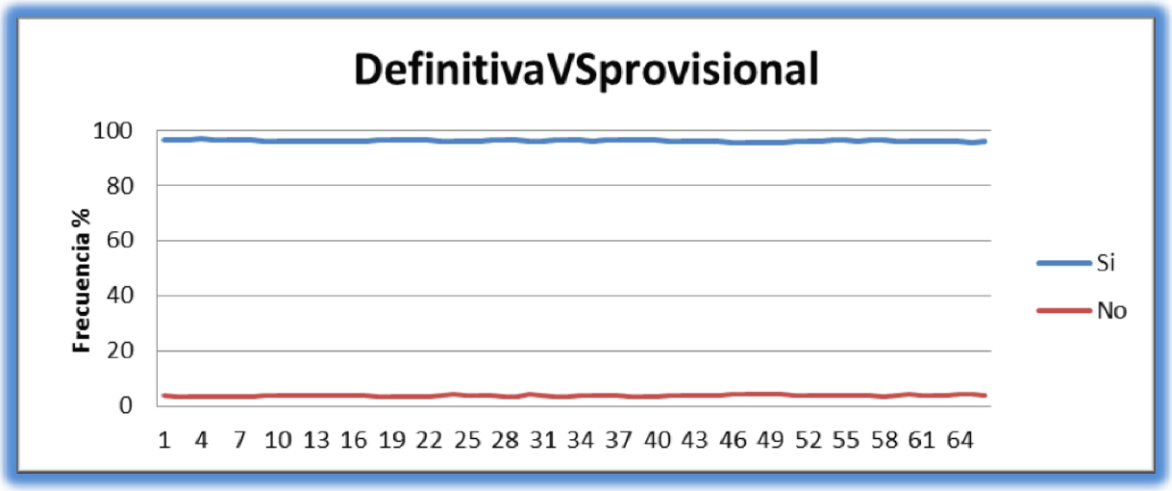

Var_indicios_aut2
Frecuencia
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoria1
Categoria2
Categoria3
Categoria4
Categoria5
Categoria6
Categoria7
Categoria8

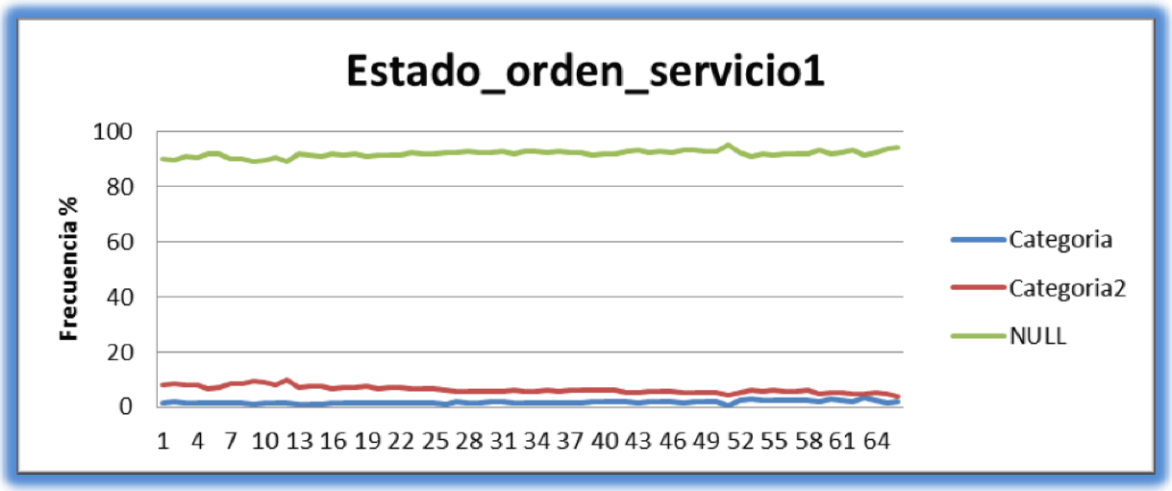

Var_indicios_aut3
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoria1
Categoria2
NULL

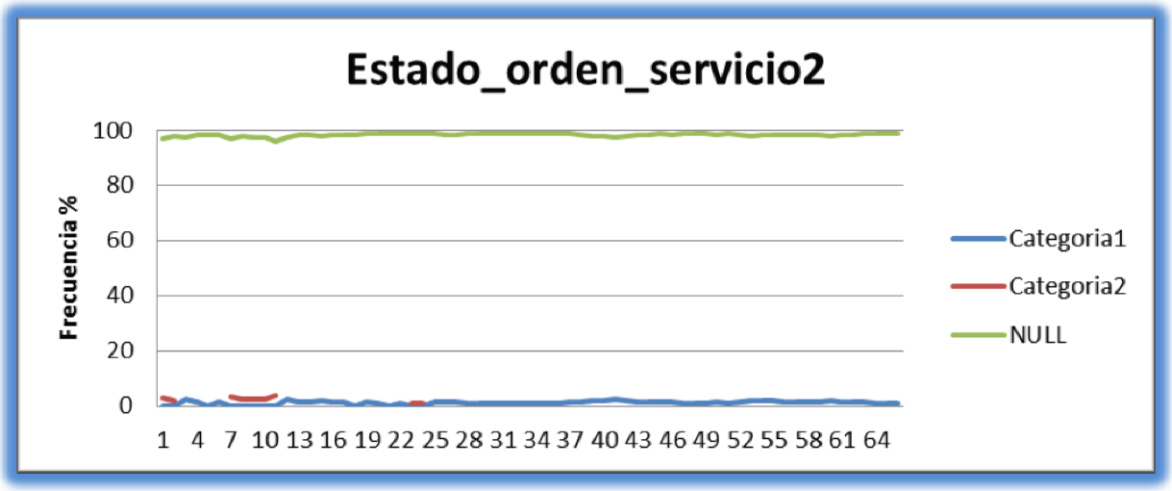

Var_indicios_manuales1
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoria1
Categoria2
Categoria3
Categoria4
NULL

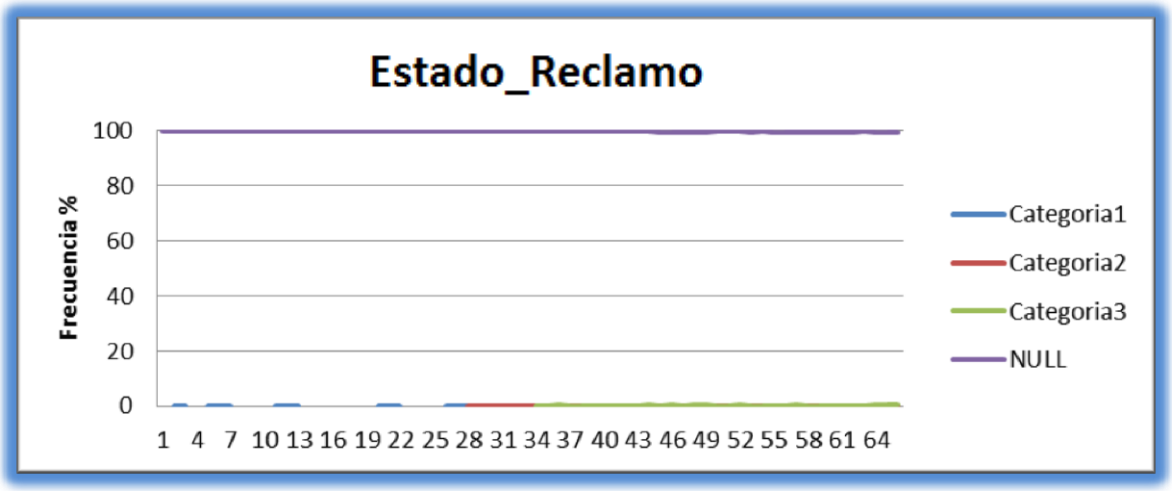

Var_indicios_manuales2
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoria1
Categoria2
Categoria3
Categoria4
Categoria5
NULL

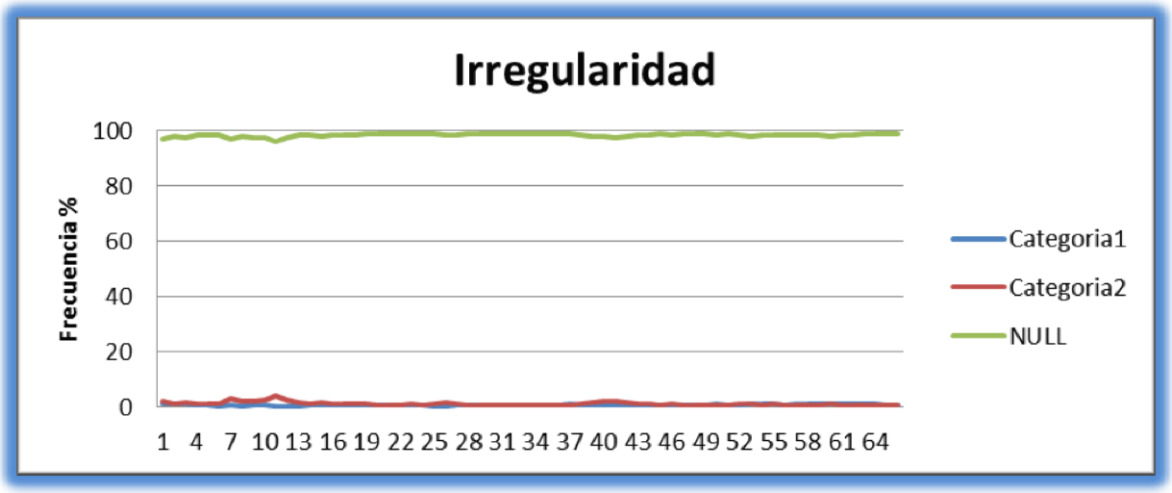

Departamento_Inspeccionador
Frecuencia
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Departamento1
Departamento2
Departamento3
Departamento4
Departamento5
Departamento6
NULL

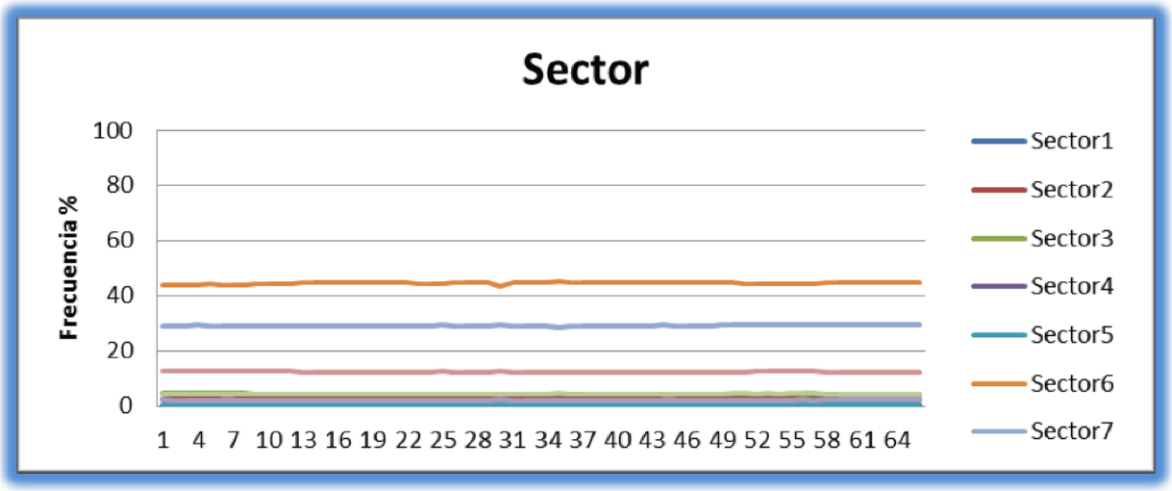

DefinitivaVSprovisional
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Si
No

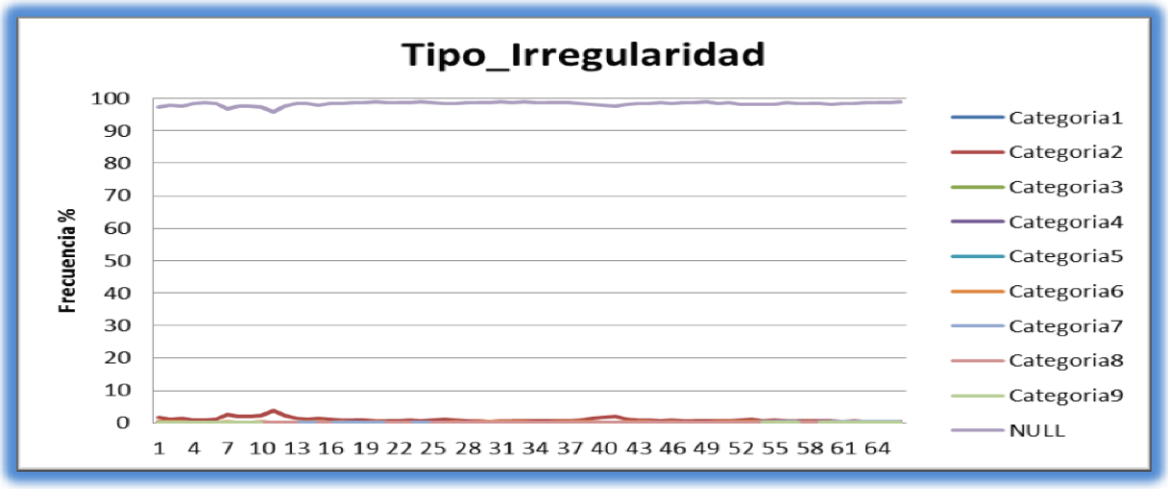

Estado_orden_servicio1
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoria
Categoria2
NULL

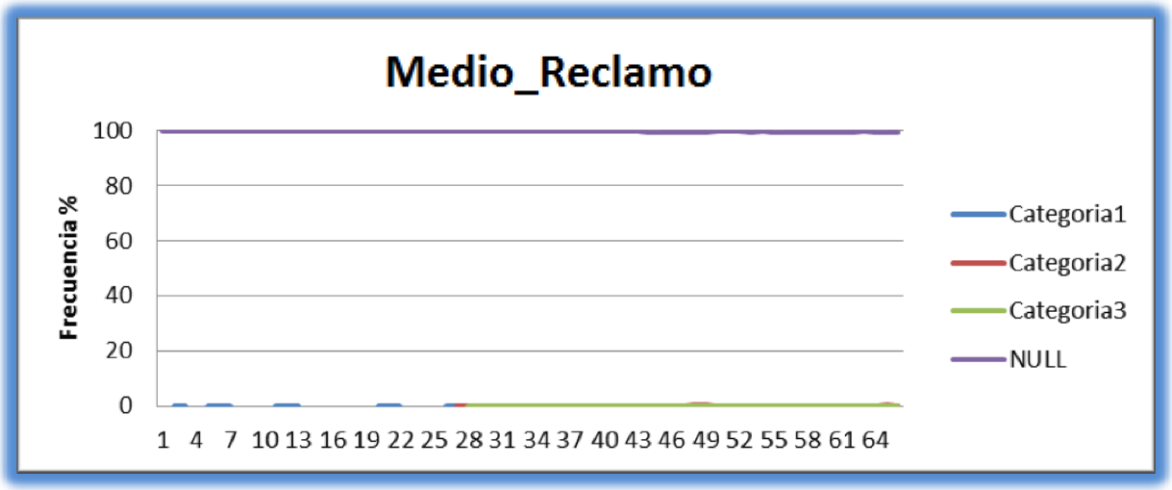

Estado_orden_servicio2
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoria1
Categoria2
NULL


Estado_Reclamo
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoria1
Categoria2
Categoria3
NULL


Irregularidad
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoria1
Categoria2
NULL


Sector
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Sector1
Sector2
Sector3
Sector4
Sector5
Sector6
Sector7


Tipo_Irregularidad
Frecuencia %
100
90
80
70
60
50
40
30
20
10
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoria1
Categoria2
Categoria3
Categoria4
Categoria5
Categoria6
Categoria7
Categoria8
Categoria9
NULL


Medio_Reclamo
Frecuencia %
100
80
60
40
20
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoria1
Categoria2
Categoria3
NULL


Tipo_Reclamo
Frecuencia
100
90
80
70
60
50
40
30
20
10
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64
Categoria1
Categoria2
Categoria3
Categoria4
Categoria5
Categoria6
Categoria7
Categoria8
Categoria9
Categoria10
NULL

11.2. Anexo Regresión Logística.

Se muestra a continuación una tabla resumen que muestra los efectos que han sido considerados significativos en el modelo de regresión logística que incluye las interacciones de orden dos, cada efecto aparecerá acompañado por su nivel de significación en el modelo, su parámetro estimado correspondiente, el error típico de dicho parámetro y del estadístico utilizado en el contraste de significación:

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-110.7	19.1923	33.2567	<.0001
CLASIFICACION_CIF6_v2	0	1	0.4992	0.1153	18.7426	<.0001
CLASIFICACION_CIF5_v2	0	1	-0.8202	0.1010	65.8918	<.0001
Tener_estimado_fraud	0	1	0.4308	0.1416	9.2527	0.0024
estm_fraude_grup_trimestre	-1	1	-2.3090	0.2094	121.6139	<.0001
estm_fraude_grup_trimestre	0	1	0.5045	0.1549	10.6116	0.0011
F1		1	1.8306	0.6090	9.0353	0.0026
F4		1	0.3874	0.1225	10.0016	0.0016
F7		1	0.2890	0.0403	51.4961	<.0001
Factor1		1	1.0227	0.0873	137.1484	<.0001
MP_CLASIFICACION_CIF		1	17.7274	0.6252	804.0342	<.0001
MP_CLASIFICACION_CIF		1	-1.7744	1.2760	1.9338	0.1643
MP_CLASIFICACION_CIF		1	59.4248	4.9809	142.3358	<.0001
MP_CLASIFICACION_CIF		1	-27.1525	15.9479	2.8987	0.0886
MP_DefinitivaVSProvi		1	-5.3752	1.5800	11.5736	0.0007
MP_Var_indicios_aut2		1	16.9586	2.7369	38.3928	<.0001
MP_Var_indicios_aut3		1	-0.2666	4.2398	0.0040	0.9499
MP_Var_indicios_aut4		1	27.0607	6.2954	18.4772	<.0001
MP_Var_indicios_manu		1	1.9119	1.7003	1.2644	0.2608
MP_estado_reclamo_re		1	180.3	52.9512	11.5903	0.0007
MP_salto_beneficio		1	-0.9484	0.1311	52.2943	<.0001
Num_trimestres_en_de		1	-0.0392	0.00469	69.7749	<.0001
antiguedad		1	1.8456	0.1444	163.3141	<.0001
indice_beneficio		1	1.6752	1.2660	1.7510	0.1858

media_CLASIFICACION_			1	225.2	50.0135	20.2668	<.0001
media_beneficio			1	-10.8967	1.8610	34.2858	<.0001
prob_local			1	10.4934	0.6754	241.3694	<.0001
sector_v2			1	-10.5275	0.9485	123.1964	<.0001
trimestre			1	0.1460	0.00971	226.0949	<.0001
variabilidad_benefic			1	19.9122	2.8261	49.6437	<.0001
MP_CLASIF*indice_ben			1	0.7218	0.1003	51.8306	<.0001
indice_be*estm_fraud	-1		1	0.0242	0.00870	7.7493	0.0054
indice_be*estm_fraud	0		1	-0.00321	0.00678	0.2244	0.6357
indice_be*Tener_esti	0		1	0.00692	0.00574	1.4556	0.2276
F4*indice_beneficio			1	-0.00958	0.0243	0.1551	0.6937
MP_CLASIF*indice_ben			1	-0.3060	0.1714	3.1876	0.0742
MP_estado*indice_ben			1	-6.8071	3.4902	3.8040	0.0511
indice_be*media_CLAS			1	-0.6609	0.1719	14.7816	0.0001
indice_be*CLASIFICAC	0		1	0.0219	0.00437	25.0681	<.0001
Factor1*indice_benef			1	-0.00100	0.00357	0.0781	0.7799
MP_Defini*prob_local			1	-0.3003	0.0556	29.1438	<.0001
MP_Var_in*indice_ben			1	0.5577	0.1847	9.1183	0.0025
Num_trime*indice_ben			1	-0.00121	0.000222	29.4977	<.0001
MP_CLASIF*prob_local			1	-4.1756	0.1346	962.9544	<.0001
MP_Var_in*MP_salto_b			1	0.5270	0.1140	21.3760	<.0001
MP_CLASIF*CLASIFICAC	0		1	0.0912	0.0232	15.4824	<.0001
MP_Var_in*prob_local			1	2.3377	0.2934	63.4606	<.0001
MP_salto_*Num_trimes			1	-0.00134	0.000218	37.6207	<.0001
MP_Defini*MP_salto_b			1	-0.0701	0.0225	9.7025	0.0018
MP_CLASIF*MP_salto_b			1	-0.1656	0.0294	31.6101	<.0001
MP_CLASIF*MP_Definit			1	-1.9013	0.0618	945.0538	<.0001
MP_CLASIF*Tener_esti	0		1	-0.5276	0.0188	785.5971	<.0001
MP_CLASIF*CLASIFICAC	0		1	-0.1197	0.0195	37.5605	<.0001
MP_Defini*Tener_esti	0		1	0.0472	0.0117	16.3456	<.0001
antigueda*Tener_esti	0		1	-0.00153	0.00500	0.0942	0.7589
MP_CLASIFI*sector_v2			1	-2.2262	0.1642	183.7746	<.0001
MP_CLASIF*antiguedad			1	-0.7818	0.0265	867.2621	<.0001
MP_CLASIF*MP_Var_ind			1	0.9419	0.3585	6.9039	0.0086
Factor1*MP_CLASIFICA			1	0.0451	0.0204	4.8776	0.0272
Tener_est*CLASIFICAC	0	0	1	0.0150	0.00497	9.1750	0.0025
antigueda*CLASIFICAC	0		1	-0.0611	0.00466	171.9961	<.0001
MP_CLASIF*MP_CLASIFI			1	10.3935	0.6211	280.0072	<.0001
media_CLA*Tener_esti	0		1	0.6383	0.2369	7.2583	0.0071
CLASIFICA*CLASIFICAC	0	0	1	-0.0111	0.00398	7.7748	0.0053

MP_CLASIF*Tener_esti	0		1	0.7881	0.1898	17.2396	<.0001
MP_CLASIF*media_CLAS			1	-9.2032	0.9069	102.9871	<.0001
F4*MP_CLASIFICACION_			1	0.5521	0.1277	18.6931	<.0001
sector_v2*CLASIFICAC	0		1	-0.4959	0.0398	155.1913	<.0001
CLASIFICA*estm_fraud	0	-1	1	-0.0163	0.0135	1.4605	0.2268
CLASIFICA*estm_fraud	0	0	1	-0.0234	0.00978	5.7000	0.0170
CLASIFICA*estm_fraud	0	-1	1	0.0922	0.0105	76.9655	<.0001
CLASIFICA*estm_fraud	0	0	1	-0.0530	0.00645	67.6185	<.0001
Factor1*MP_Definitiv			1	0.0522	0.00483	116.7800	<.0001
media_CLA*CLASIFICAC	0		1	0.7492	0.1955	14.6803	0.0001
sector_v2*Tener_esti	0		1	0.1000	0.0348	8.2491	0.0041
MP_CLASIF*CLASIFICAC	0		1	2.1046	0.1464	206.5289	<.0001
antiguedad*sector_v2			1	0.5845	0.0381	235.9142	<.0001
Factor1*CLASIFICACIO	0		1	0.0265	0.00320	68.8995	<.0001
sector_v2*CLASIFICAC	0		1	0.2026	0.0342	35.0949	<.0001
MP_Defini*MP_estado_			1	22.6647	4.2692	28.1846	<.0001
trimestre*Tener_esti	0		1	0.00426	0.000340	156.8460	<.0001
trimestre*CLASIFICAC	0		1	0.00352	0.000304	133.9398	<.0001
MP_CLASIF*CLASIFICAC	0		1	-0.8454	0.1504	31.5874	<.0001
F4*Tener_estimado_fr	0		1	0.0510	0.0201	6.4465	0.0111
Factor1*media_CLASIF			1	-0.7536	0.1551	23.6153	<.0001
Factor1*CLASIFICACIO	0		1	-0.0106	0.00316	11.1654	0.0008
antigueda*estm_fraud	-1		1	0.0706	0.0104	46.2445	<.0001
antigueda*estm_fraud	0		1	0.0369	0.00721	26.1718	<.0001
MP_CLASIF*MP_Var_ind			1	13.6560	1.8667	53.5162	<.0001
antigueda*media_CLAS			1	0.6955	0.2243	9.6135	0.0019
MP_CLASIFI*sector_v2			1	14.4182	1.9311	55.7470	<.0001
MP_CLASIFI*trimestre			1	0.00488	0.00155	9.8615	0.0017
MP_CLASIF*MP_Var_ind			1	-12.3425	2.0602	35.8923	<.0001
MP_Var_ind*trimestre			1	-0.0563	0.00485	134.7366	<.0001
MP_Var_in*prob_local			1	3.6471	1.7096	4.5508	0.0329
MP_CLASIF*MP_Definit			1	-0.0534	0.1191	0.2012	0.6538
media_CLAS*trimestre			1	-0.0821	0.0137	36.2035	<.0001
MP_CLASIF*CLASIFICAC	0		1	-0.5874	0.1053	31.1014	<.0001
MP_CLASIF*MP_Definit			1	-2.9624	0.8342	12.6104	0.0004
MP_CLASIF*antiguedad			1	-3.3115	0.1935	292.8714	<.0001
antiguedad*trimestre			1	0.00331	0.000386	73.8537	<.0001
MP_Definit*trimestre			1	-0.00672	0.000878	58.5296	<.0001
MP_Defini*media_CLAS			1	-0.4420	0.4604	0.9218	0.3370
Factor1*MP_CLASIFICA			1	-1.0539	0.1073	96.5452	<.0001

MP_Defini*estm_fraud	-1	1	-0.1001	0.0138	52.4474	<.0001
MP_Defini*estm_fraud	0	1	0.0303	0.0103	8.5603	0.0034
MP_CLASIFI*trimestre		1	-0.2905	0.00926	985.5509	<.0001
media_CLA*estm_fraud	-1	1	3.5287	0.4728	55.7071	<.0001
media_CLA*estm_fraud	0	1	-1.7163	0.3619	22.4946	<.0001
sector_v2*estm_fraud	-1	1	0.2733	0.0705	15.0149	0.0001
sector_v2*estm_fraud	0	1	0.00154	0.0538	0.0008	0.9771
MP_CLASIF*CLASIFICAC	0	1	0.4793	0.1013	22.3949	<.0001
MP_Defini*MP_Var_ind		1	0.3086	0.0556	30.8214	<.0001
MP_Defini*MP_Var_ind		1	-3.4447	0.3585	92.3351	<.0001
F1*MP_CLASIFICACION_		1	0.1053	0.0168	39.1163	<.0001
MP_estado*media_CLAS		1	-586.0	138.2	17.9715	<.0001
Factor1*estm_fraude_	-1	1	-0.0321	0.00998	10.3092	0.0013
Factor1*estm_fraude_	0	1	0.0450	0.00616	53.3459	<.0001
MP_CLASIF*prob_local		1	-1.1640	0.2896	16.1522	<.0001
MP_CLASIF*MP_Var_ind		1	-9.0389	1.7373	27.0707	<.0001
MP_Var_in*CLASIFICAC	0	1	1.7330	0.3318	27.2743	<.0001
MP_CLASIF*antiguedad		1	-1.8574	0.1331	194.7454	<.0001
MP_Var_in*Tener_esti	0	1	-0.8327	0.0605	189.1821	<.0001
MP_CLASIFI*trimestre		1	-0.0165	0.00739	4.9890	0.0255
MP_Var_in*antiguedad		1	-1.5518	0.3784	16.8216	<.0001
Factor1*MP_CLASIFICA		1	-0.2327	0.0801	8.4299	0.0037
MP_Defini*MP_Var_ind		1	3.0447	0.3475	76.7651	<.0001
MP_Var_in*MP_Var_ind		1	-7.7387	1.2744	36.8747	<.0001
MP_Var_in*CLASIFICAC	0	1	-0.8853	0.2457	12.9815	0.0003
MP_CLASIF*media_CLAS		1	-21.3767	3.4824	37.6820	<.0001
F7*prob_local		1	-0.2432	0.0344	50.0996	<.0001
Factor1*trimestre		1	-0.00489	0.000239	418.1012	<.0001
MP_CLASIF*MP_CLASIFI		1	-4.1495	0.2465	283.4121	<.0001
sector_v2*trimestre		1	-0.0425	0.00258	271.3327	<.0001
MP_Var_in*prob_local		1	-5.7651	1.2645	20.7867	<.0001
MP_salto_b*trimestre		1	-0.00218	0.000414	27.7190	<.0001
F1*MP_Var_indicios_m		1	0.3041	0.0592	26.4281	<.0001
F7*MP_salto_benefici		1	0.0378	0.00829	20.7800	<.0001
F7*MP_DefinitivaVSPr		1	0.0742	0.0128	33.4474	<.0001
MP_CLASIF*MP_estado_		1	141.2	44.0183	10.2915	0.0013
MP_Var_in*Tener_esti	0	1	-0.9987	0.1714	33.9421	<.0001
trimestre*estm_fraud	-1	1	0.00266	0.000694	14.7188	0.0001
trimestre*estm_fraud	0	1	-0.00539	0.000515	109.5308	<.0001
MP_Var_ind*trimestre		1	-0.0707	0.0232	9.3068	0.0023

Factor1*MP_Var_indic		1	-0.3342	0.1275	6.8754	0.0087
MP_Var_in*media_CLAS		1	11.3943	2.4922	20.9037	<.0001
F1*MP_DefinitivaVSPr		1	-0.0265	0.0104	6.4872	0.0109
MP_Var_ind*trimestre		1	0.1959	0.0302	42.1223	<.0001
F7*MP_CLASIFICACION_		1	-0.3932	0.0332	140.0426	<.0001
MP_CLASIF*CLASIFICAC	0	1	0.1320	0.0481	7.5232	0.0061
MP_Var_ind*trimestre		1	-0.1216	0.0231	27.6824	<.0001
MP_CLASIF*Num_trimes		1	0.0525	0.00542	93.7316	<.0001
Num_trime*antiguedad		1	0.00105	0.000247	17.8688	<.0001
MP_CLASIF*MP_Definit		1	-0.6959	0.1233	31.8541	<.0001
Num_trime*media_CLAS		1	0.0370	0.00717	26.6050	<.0001
MP_CLASIF*MP_Var_ind		1	-14.4686	3.7480	14.9024	0.0001
MP_CLASIF*MP_Var_ind		1	-36.1497	11.8742	9.2683	0.0023
MP_Var_in*media_CLAS		1	-38.4670	6.2946	37.3463	<.0001
MP_CLASIF*CLASIFICAC	0	1	0.2677	0.0426	39.5577	<.0001
MP_CLASIF*MP_Var_ind		1	5.9827	0.6562	83.1139	<.0001
MP_CLASIF*antiguedad		1	0.1128	0.0771	2.1386	0.1436
MP_Var_in*antiguedad		1	1.7974	0.3552	25.5980	<.0001
MP_CLASIF*Tener_esti	0	1	0.1280	0.0437	8.5640	0.0034
MP_Var_ind*sector_v2		1	6.6580	1.3267	25.1848	<.0001
Factor1*Num_trimestr		1	0.00188	0.000153	151.4502	<.0001
MP_CLASIFI*sector_v2		1	-1.1845	0.4581	6.6861	0.0097
Num_trimes*sector_v2		1	-0.0100	0.00151	44.4870	<.0001
MP_Var_in*MP_Var_ind		1	-14.5213	2.0964	47.9788	<.0001
F1*Tener_estimado_fr	0	1	0.00900	0.00320	7.8788	0.0050
MP_salto_b*sector_v2		1	-0.4000	0.0441	82.3962	<.0001
F7*Tener_estimado_fr	0	1	0.0503	0.00620	65.6806	<.0001
F4*MP_CLASIFICACION_		1	-1.3224	0.2736	23.3533	<.0001
MP_CLASIFI*trimestre		1	-0.0246	0.00344	51.1375	<.0001
MP_CLASIF*MP_CLASIFI		1	10.6705	1.3967	58.3678	<.0001
F7*CLASIFICACION_CIF	0	1	-0.0324	0.00558	33.7217	<.0001
F7*MP_Var_indicios_m		1	0.4038	0.0650	38.5428	<.0001
F7*trimestre		1	-0.00239	0.000401	35.5855	<.0001
MP_Var_in*Num_trimes		1	0.0223	0.00493	20.4972	<.0001
F7*CLASIFICACION_CIF	0	1	-0.0297	0.00519	32.7201	<.0001
F1*MP_estado_reclamo		1	-4.7455	1.6658	8.1152	0.0044
F1*antiguedad		1	-0.0216	0.00515	17.6485	<.0001
F1*media_CLASIFICACI		1	-0.5278	0.2006	6.9201	0.0085
MP_Var_in*MP_salto_b		1	1.8903	0.2289	68.2124	<.0001
antigueda*prob_local		1	-0.6299	0.0277	515.8205	<.0001

F1*MP_CLASIFICACION_		1	0.2281	0.0450	25.6445	<.0001
F7*Num_trimestres_en		1	0.000339	0.000181	3.5168	0.0608
MP_CLASIF*MP_Var_ind		1	-2.6556	1.8545	2.0504	0.1522
media_CLA*prob_local		1	4.3144	1.0315	17.4936	<.0001
F1*MP_salto_benefici		1	0.0206	0.00488	17.8003	<.0001
F1*estm_fraude_grup_	-1	1	-0.0154	0.00646	5.6438	0.0175
F1*estm_fraude_grup_	0	1	-0.00298	0.00489	0.3711	0.5424
F7*estm_fraude_grup_	-1	1	0.0577	0.0120	22.9390	<.0001
F7*estm_fraude_grup_	0	1	-0.0347	0.00907	14.6581	0.0001
MP_CLASIF*Num_trimes		1	-0.0193	0.00234	67.8271	<.0001
F1*trimestre		1	0.00157	0.000282	30.8743	<.0001
prob_loca*estm_fraud	-1	1	0.2187	0.0714	9.3983	0.0022
prob_loca*estm_fraud	0	1	-0.3023	0.0423	50.9774	<.0001
prob_loca*CLASIFICAC	0	1	0.1020	0.0236	18.6004	<.0001
prob_loca*CLASIFICAC	0	1	-0.0821	0.0225	13.2489	0.0003
MP_CLASIF*prob_local		1	-1.7425	0.5488	10.0810	0.0015
prob_local*sector_v2		1	-2.0132	0.2255	79.6654	<.0001
MP_CLASIF*MP_salto_b		1	1.0434	0.1585	43.3578	<.0001
MP_salto_*estm_fraud	-1	1	-0.0349	0.0104	11.2625	0.0008
MP_salto_*estm_fraud	0	1	-0.00709	0.00697	1.0359	0.3088
F7*indice_beneficio		1	0.000468	0.00654	0.0051	0.9430
F1*indice_beneficio		1	-0.00952	0.00416	5.2392	0.0221
MP_CLASIF*indice_ben		1	0.2059	0.0682	9.1087	0.0025
indice_ben*trimestre		1	0.00282	0.000414	46.5391	<.0001
prob_local*trimestre		1	0.0355	0.00174	416.6728	<.0001
MP_Defini*indice_ben		1	-0.0750	0.0139	29.2938	<.0001
indice_be*prob_local		1	0.2398	0.0296	65.6137	<.0001
MP_Var_in*indice_ben		1	0.8764	0.0812	116.3840	<.0001
F1*media_beneficio		1	0.0362	0.00610	35.2451	<.0001
F4*media_beneficio		1	-0.0798	0.0293	7.4438	0.0064
F7*media_beneficio		1	-0.0335	0.0107	9.8553	0.0017
F7*variabilidad_bene		1	0.0570	0.0132	18.6836	<.0001
MP_CLASIF*variabilid		1	-1.0907	0.2060	28.0326	<.0001
MP_estado*media_bene		1	28.6819	5.1497	31.0207	<.0001
MP_estado*variabilid		1	-46.4306	7.8313	35.1515	<.0001
indice_be*media_bene		1	0.00594	0.00143	17.2716	<.0001
indice_be*variabilid		1	0.0162	0.00368	19.2681	<.0001
media_ben*variabilid		1	-0.0519	0.00432	144.1490	<.0001
variabili*CLASIFICAC	0	1	-0.0262	0.00885	8.7792	0.0030
media_ben*Tener_esti	0	1	-0.0215	0.00661	10.6057	0.0011

variabili*Tener_esti	0	1	-0.0810	0.0114	50.5201	<.0001
MP_CLASIF*media_bene		1	-0.1043	0.0335	9.6757	0.0019
MP_CLASIF*variabilid		1	-0.7739	0.0584	175.7345	<.0001
MP_CLASIF*media_bene		1	0.4662	0.0842	30.6756	<.0001
MP_CLASIF*variabilid		1	-0.5847	0.1214	23.1790	<.0001
variabili*estm_fraud	-1	1	-0.1436	0.0214	45.1921	<.0001
variabili*estm_fraud	0	1	0.0283	0.0157	3.2447	0.0717
MP_Defini*media_bene		1	0.0264	0.0221	1.4279	0.2321
MP_Var_in*media_bene		1	2.4287	0.2802	75.1225	<.0001
MP_Var_in*variabilid		1	2.0841	0.5382	14.9948	0.0001
MP_Var_in*variabilid		1	-5.7267	0.6130	87.2897	<.0001
MP_Var_in*media_bene		1	-0.2388	0.1259	3.5989	0.0578
MP_Var_in*variabilid		1	-0.8514	0.0927	84.3144	<.0001
MP_salto_*media_bene		1	0.0693	0.00710	95.2343	<.0001
antigueda*variabilid		1	-0.0284	0.00883	10.3171	0.0013
media_ben*prob_local		1	-0.0257	0.0355	0.5221	0.4699
media_bene*sector_v2		1	-0.3221	0.0539	35.6685	<.0001
media_bene*trimestre		1	-0.00149	0.000588	6.3905	0.0115
prob_loca*variabilid		1	-0.7764	0.0606	164.3564	<.0001
sector_v2*variabilid		1	0.2944	0.0885	11.0572	0.0009
trimestre*variabilid		1	0.00224	0.000788	8.1039	0.0044

11.3. Anexo Árbol de Random Forest.

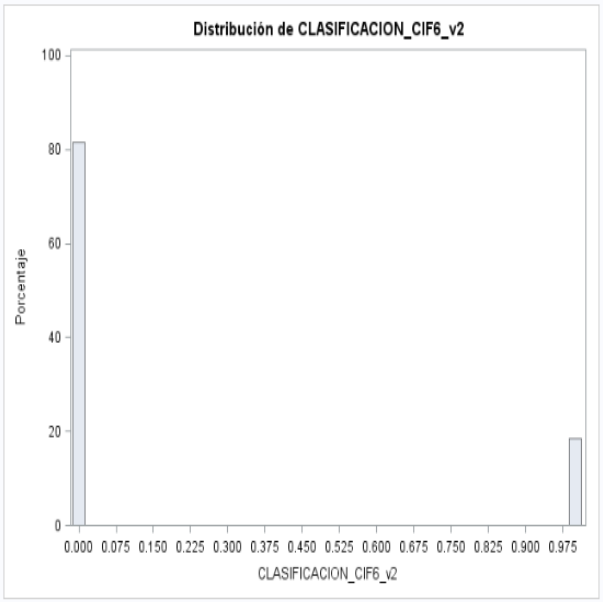
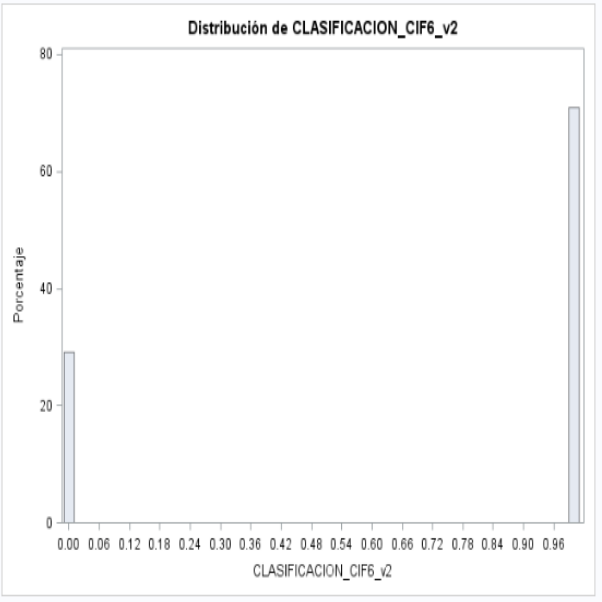
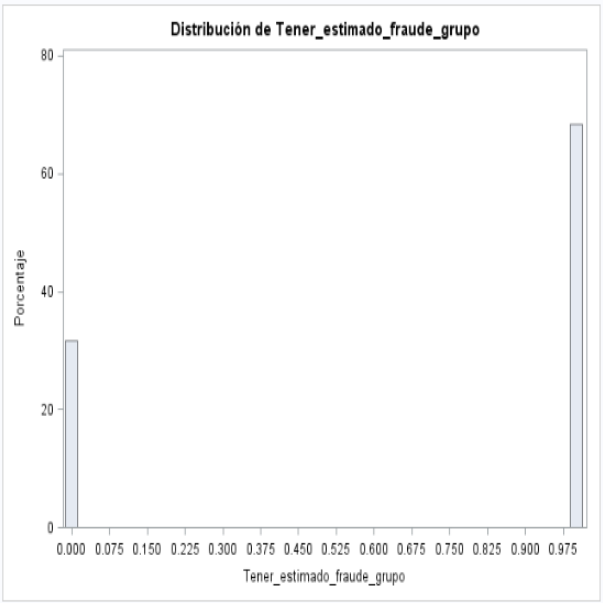
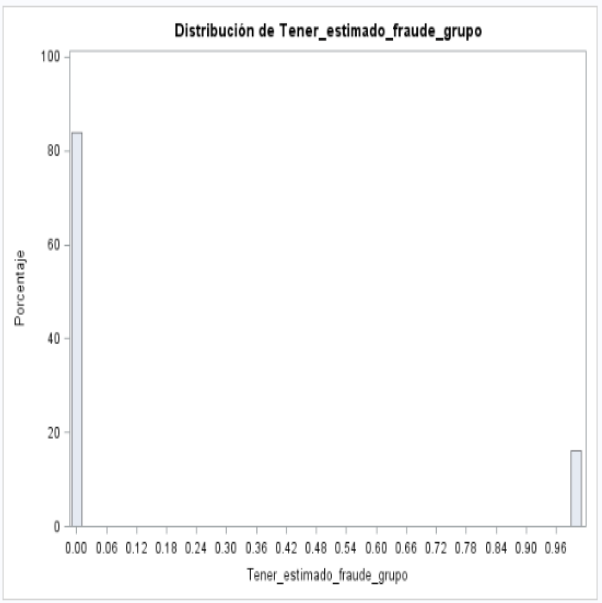
Variables	Importancia
MP_CLASIFICACION_CIF3	0,728031585
prob_local	1
Tener_estimado_fraude_grupo	0,234066554
MP_DefinitivaVSProvisional	0,553186689
CLASIFICACION_CIF5_v2	0,195262267
MP_Var_indicios_manuales1	0,411280316
MP_CLASIFICACION_CIF8	0,282233503
MP_var_indicios_manuales2	0,434179357
trimestre	0,918330513

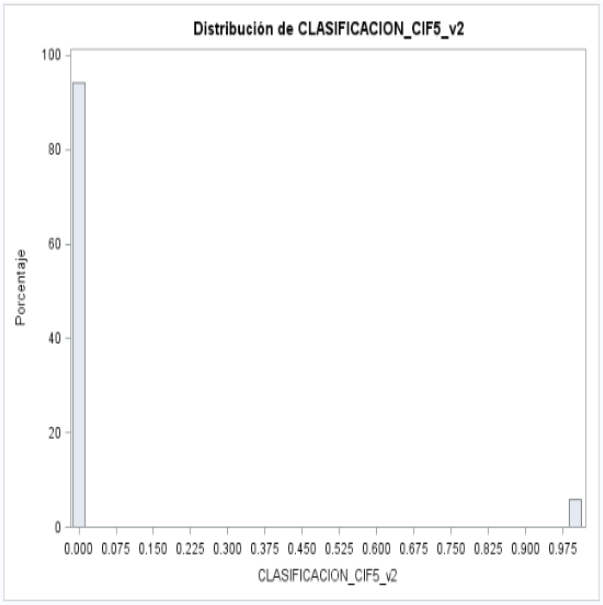
Factor1	0,480654258
sector_v2	0,45572476
F7	0,53322053
Num_trimestres_en_deuda	0,442752397
MP_CLASIFICACION_CIF7	0,500507614
CLASIFICACION_CIF6_v2	0,276029329
variabilidad_beneficio	0,573378455
estm_fraude_grup_trim_v2	0,322278624
MP_CLASIFICACION_CIF9	0,289904117
media_CLASIFICACION_CIF1	0,059334461
media_beneficio	0,402820079
MP_Var_indicios_aut3	0,256514382
indice_beneficio	0,356909193
MP_salto_beneficio	0,341342358
F1	0,188494078
F6	0,075578116
MP_estado_reclamo_refact	0,058657642
F4	0,012633954
MP_medio_reclamo_refact	0,026847152
F2	0,010829103
F8	0,000225606
MP_Var_indicios_aut4	0,237901861
antiguedad	0,127580372
MP_Var_indicios_aut2	0,22357586

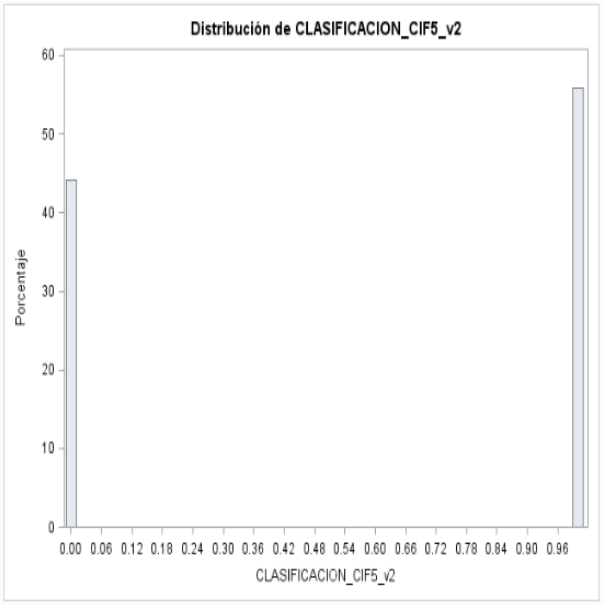
11.4. Anexo Árbol de Gradient Boosting.

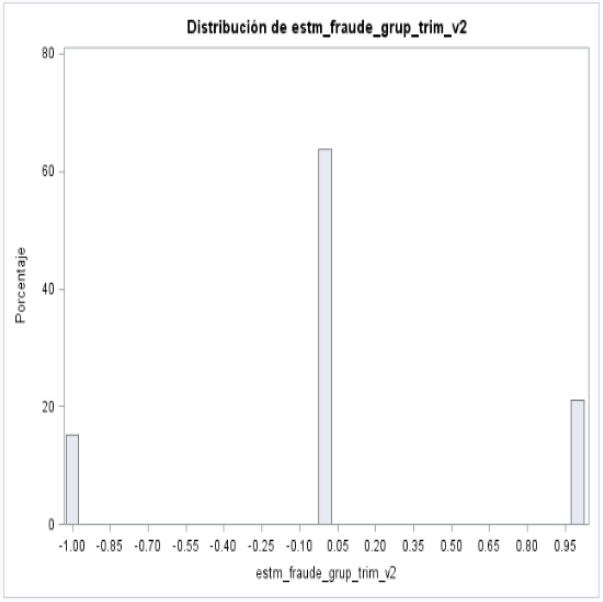
VAR	IMPORTANCIA
MP_CLASIFICACION_CIF3	1
prob_local	0,80585421
trimestre	0,486459452
variabilidad_beneficio	0,432142784
media_beneficio	0,416506036
MP_DefinitivaVSProvisional	0,378369888
	0,366603943
indice_beneficio	0,346465325
Tener_estimado_fraude_grupo	0,33031288
MP_CLASIFICACION_CIF8	0,327768454
F7	0,320665835
MP_CLASIFICACION_CIF7	0,306753982
MP_salto_beneficio	0,305866592
Num_trimestres_en_deuda	0,299764361
MP_Var_indicios_manuales1	0,282113475
MP_var_indicios_manuales2	0,273277984
sector_v2	0,233018788
MP_CLASIFICACION_CIF9	0,21905813
F1	0,208953136
MP_Var_indicios_aut3	0,196392874
MP_Var_indicios_aut4	0,184056672
MP_Var_indicios_aut2	0,162967539
estm_fraude_grup_trim_v2	0,157946489
MP_medio_reclamo_refact	0,128515686
CLASIFICACION_CIF6_v2	0,122993628
F6	0,112636262
CLASIFICACION_CIF5_v2	0,109772003
MP_estado_reclamo_refact	0,084102801
F4	0,053369736
media_CLASIFICACION_CIF1	0,041197352
F2	0,040945769
antiguedad	0,034364968
F8	0

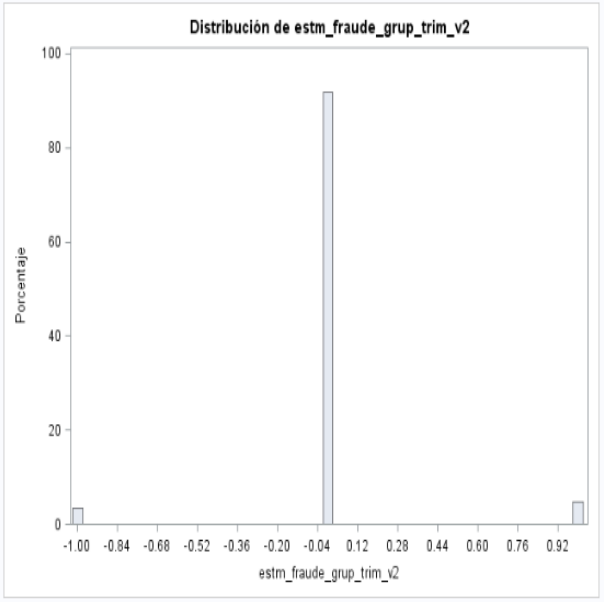
11.5. Post-Análisis.

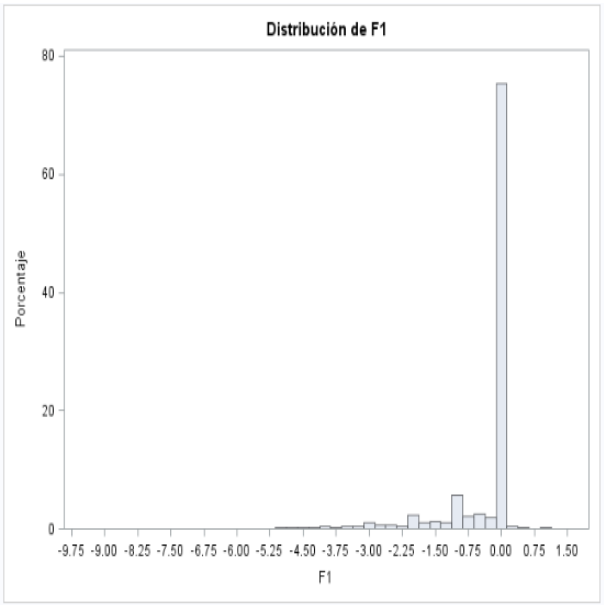
Fraude_Estimado=-1	Fraude_Estimado=1
	
	

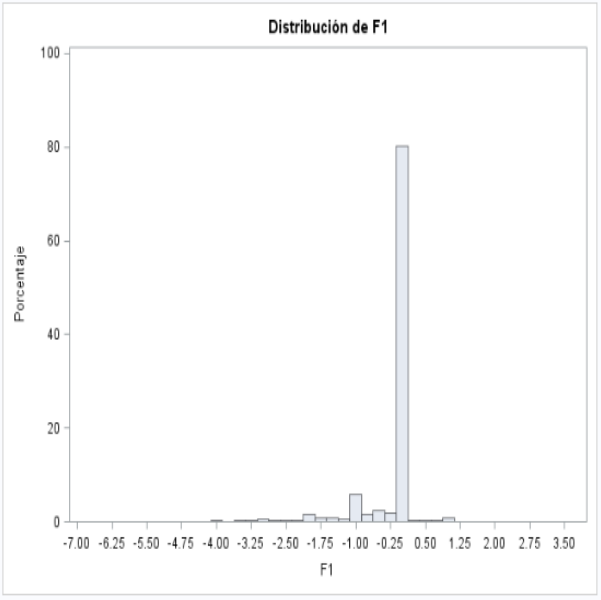

Distribución de CLASIFICACION_CIF5_v2
Porcentaje
100
80
60
40
20
0
0.000 0.075 0.150 0.225 0.300 0.375 0.450 0.525 0.600 0.675 0.750 0.825 0.900 0.975
CLASIFICACION_CIF5_v2

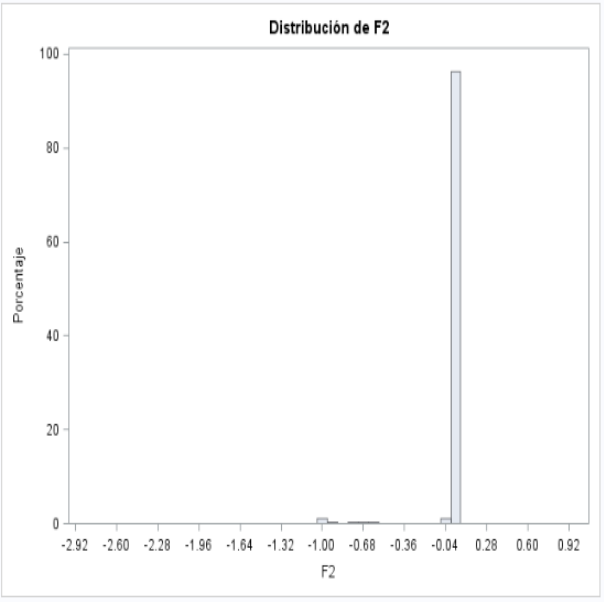

Distribución de CLASIFICACION_CIF5_v2
Porcentaje
60
50
40
30
20
10
0
0.00 0.06 0.12 0.18 0.24 0.30 0.36 0.42 0.48 0.54 0.60 0.66 0.72 0.78 0.84 0.90 0.96
CLASIFICACION_CIF5_v2

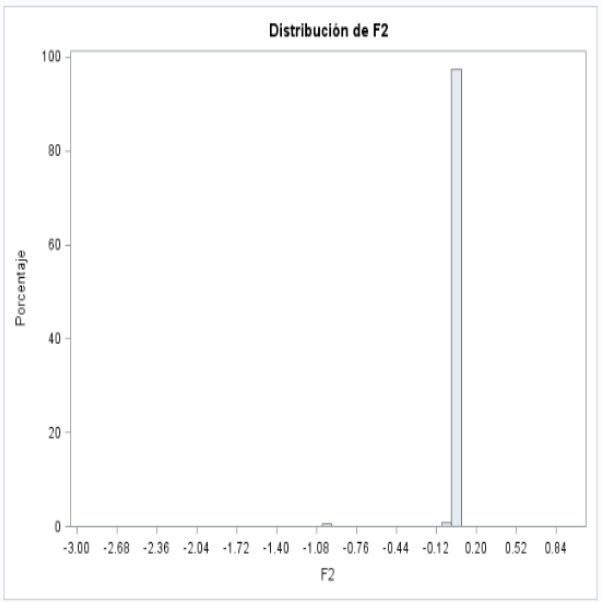

Distribución de estm_fraude_grup_trim_v2
Porcentaje
80
60
40
20
0
-1.00 -0.85 -0.70 -0.55 -0.40 -0.25 -0.10 0.05 0.20 0.35 0.50 0.65 0.80 0.95
estm_fraude_grup_trim_v2

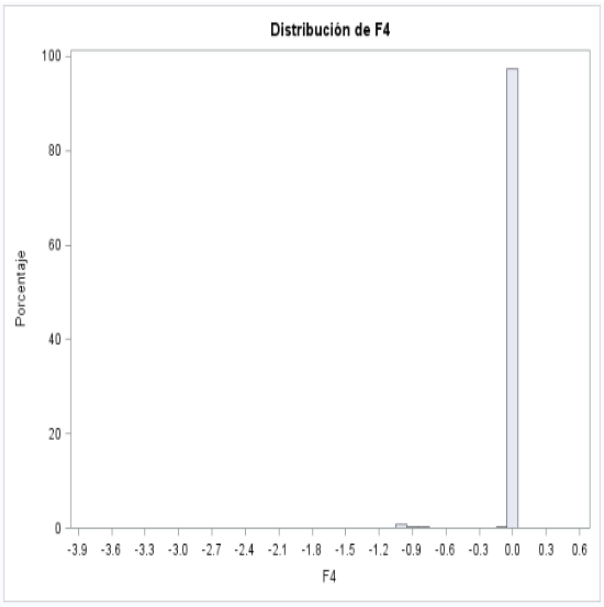

Distribución de estm_fraude_grup_trim_v2
Porcentaje
100
80
60
40
20
0
-1.00 -0.84 -0.68 -0.52 -0.36 -0.20 -0.04 0.12 0.28 0.44 0.60 0.76 0.92
estm_fraude_grup_trim_v2

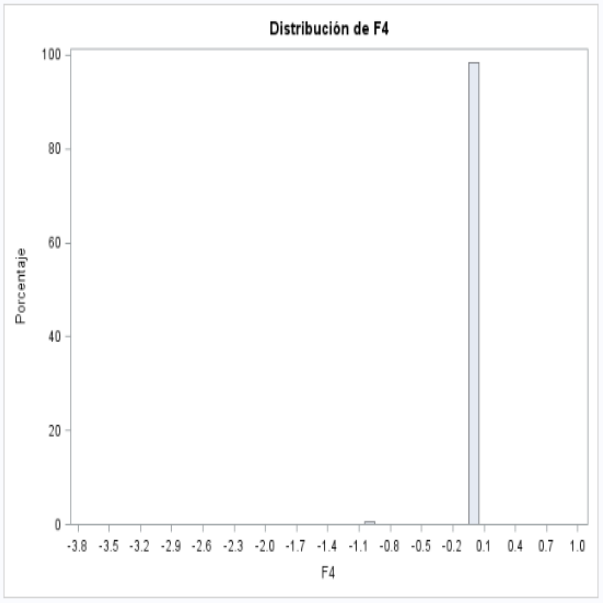

Distribución de F1
Porcentaje
0
20
40
60
80
-9.75 -9.00 -8.25 -7.50 -6.75 -6.00 -5.25 -4.50 -3.75 -3.00 -2.25 -1.50 -0.75 0.00 0.75 1.50
F1

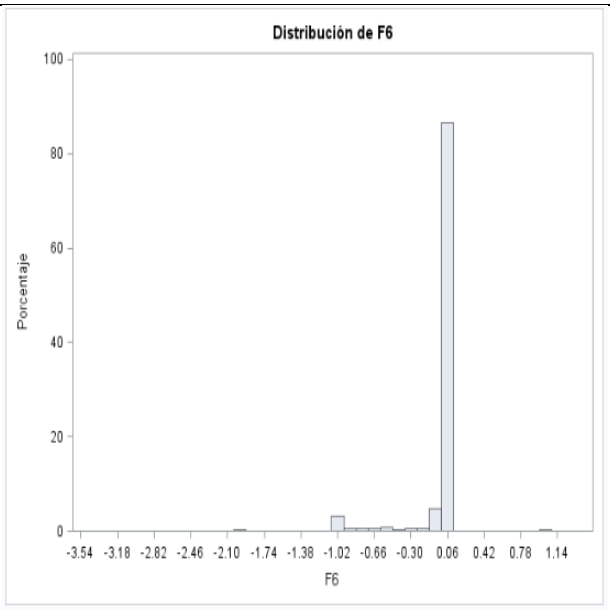

Distribución de F1
Porcentaje
0
20
40
60
80
100
-7.00 -6.25 -5.50 -4.75 -4.00 -3.25 -2.50 -1.75 -1.00 -0.25 0.50 1.25 2.00 2.75 3.50
F1

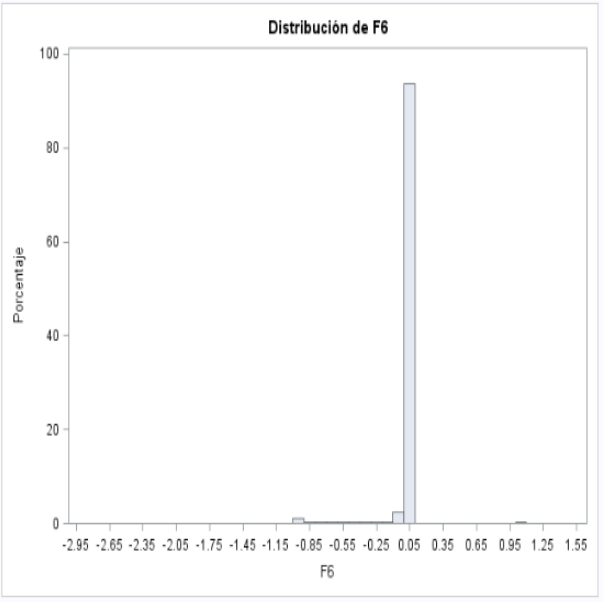

Distribución de F2
Porcentaje
0
20
40
60
80
100
-2.92 -2.60 -2.28 -1.96 -1.64 -1.32 -1.00 -0.68 -0.36 -0.04 0.28 0.60 0.92
F2

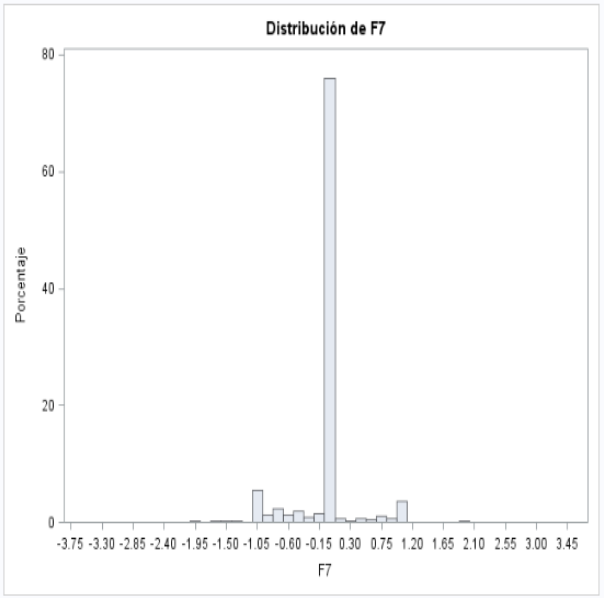

Distribución de F2
Porcentaje
0
20
40
60
80
100
-3.00 -2.68 -2.36 -2.04 -1.72 -1.40 -1.08 -0.76 -0.44 -0.12 0.20 0.52 0.84
F2

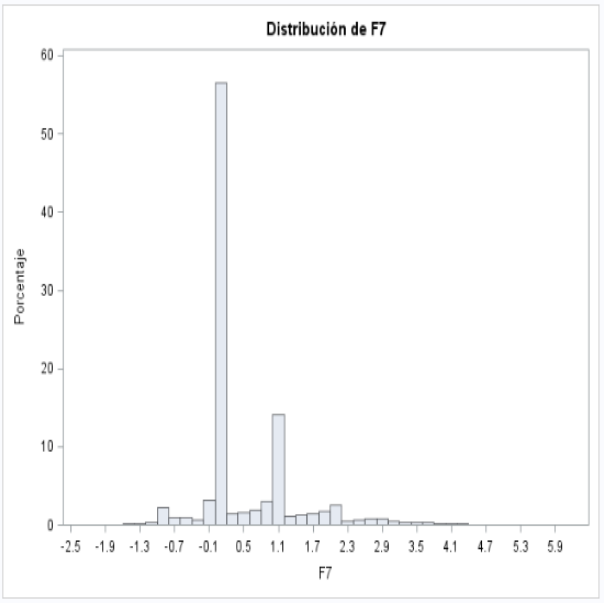

Distribución de F4
Porcentaje
100
80
60
40
20
0
-3.9 -3.6 -3.3 -3.0 -2.7 -2.4 -2.1 -1.8 -1.5 -1.2 -0.9 -0.6 -0.3 0.0 0.3 0.6
F4

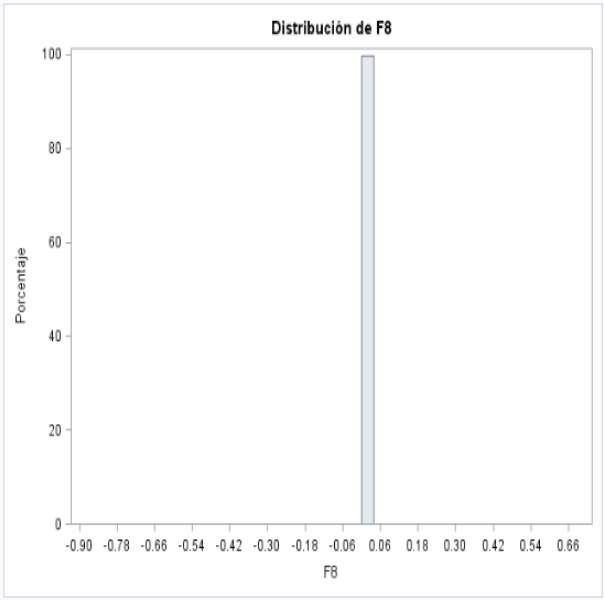

Distribución de F4
Porcentaje
100
80
60
40
20
0
-3.8 -3.5 -3.2 -2.9 -2.6 -2.3 -2.0 -1.7 -1.4 -1.1 -0.8 -0.5 -0.2 0.1 0.4 0.7 1.0
F4

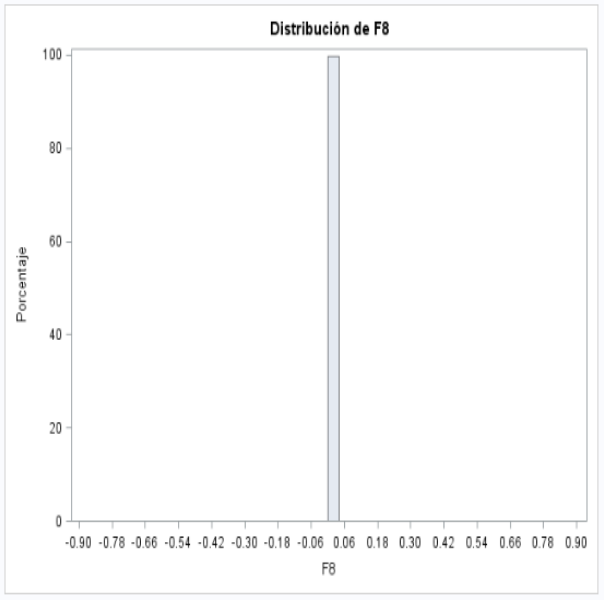

Distribución de F6
Porcentaje
100
80
60
40
20
0
-3.54 -3.18 -2.82 -2.46 -2.10 -1.74 -1.38 -1.02 -0.66 -0.30 0.06 0.42 0.78 1.14
F6

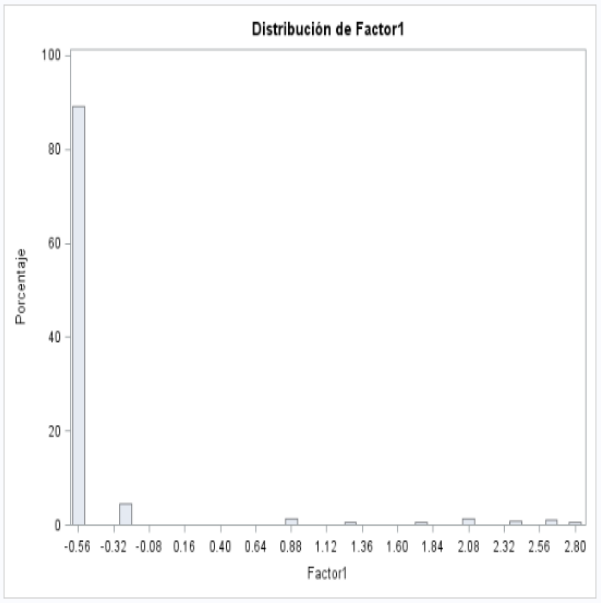

Distribución de F6
Porcentaje
100
80
60
40
20
0
-2.95 -2.65 -2.35 -2.05 -1.75 -1.45 -1.15 -0.85 -0.55 -0.25 0.05 0.35 0.65 0.95 1.25 1.55
F6

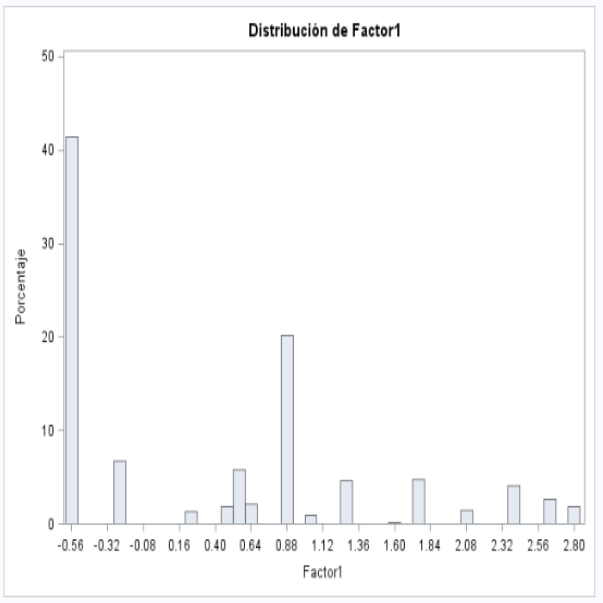

Distribución de F7
Porcentaje
0
20
40
60
80
-3.75 -3.30 -2.85 -2.40 -1.95 -1.50 -1.05 -0.60 -0.15 0.30 0.75 1.20 1.65 2.10 2.55 3.00 3.45
F7

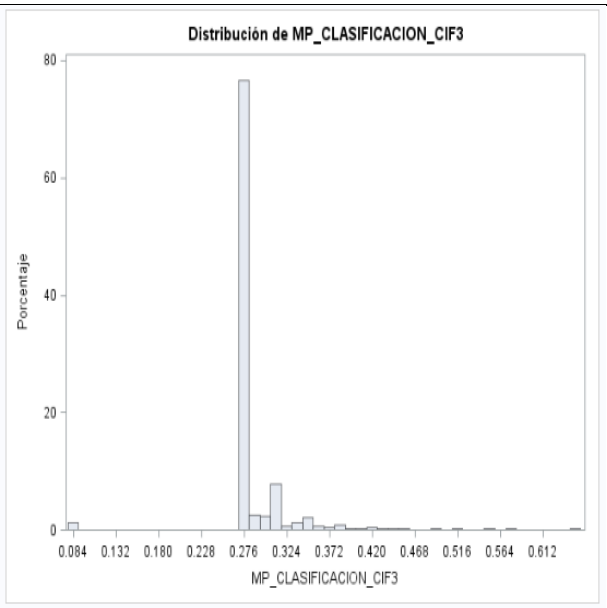

Distribución de F7
Porcentaje
0
10
20
30
40
50
60
-2.5 -1.9 -1.3 -0.7 -0.1 0.5 1.1 1.7 2.3 2.9 3.5 4.1 4.7 5.3 5.9
F7

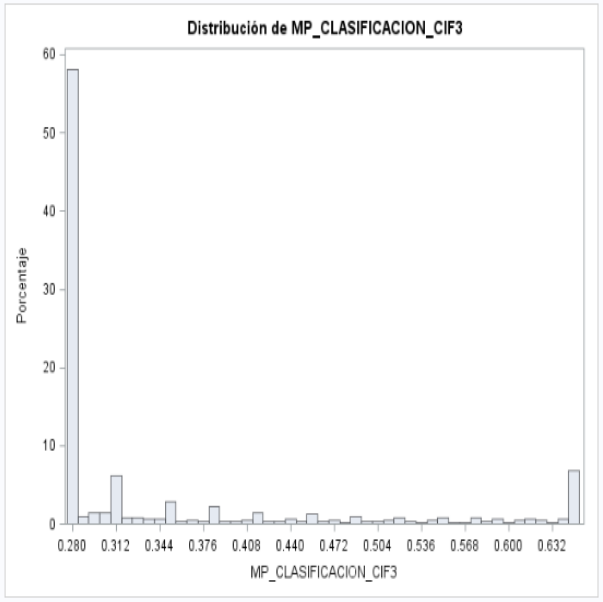

Distribución de F8
Porcentaje
0
20
40
60
80
100
-0.90 -0.78 -0.66 -0.54 -0.42 -0.30 -0.18 -0.06 0.06 0.18 0.30 0.42 0.54 0.66
F8

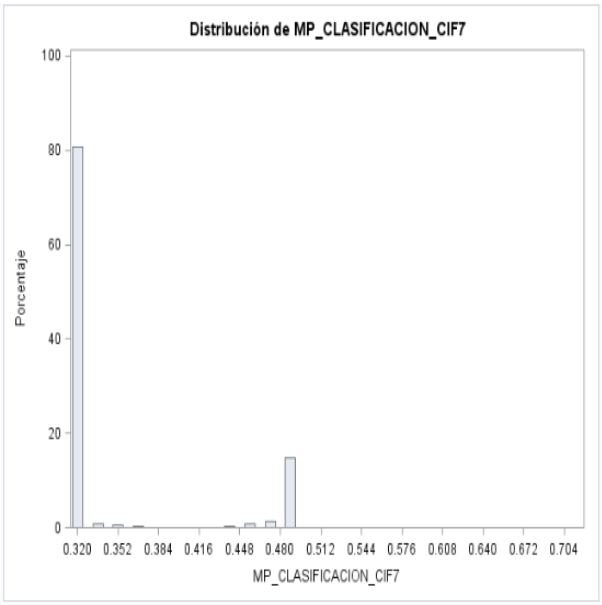

Distribución de F8
Porcentaje
0
20
40
60
80
100
-0.90 -0.78 -0.66 -0.54 -0.42 -0.30 -0.18 -0.06 0.06 0.18 0.30 0.42 0.54 0.66 0.78 0.90
F8

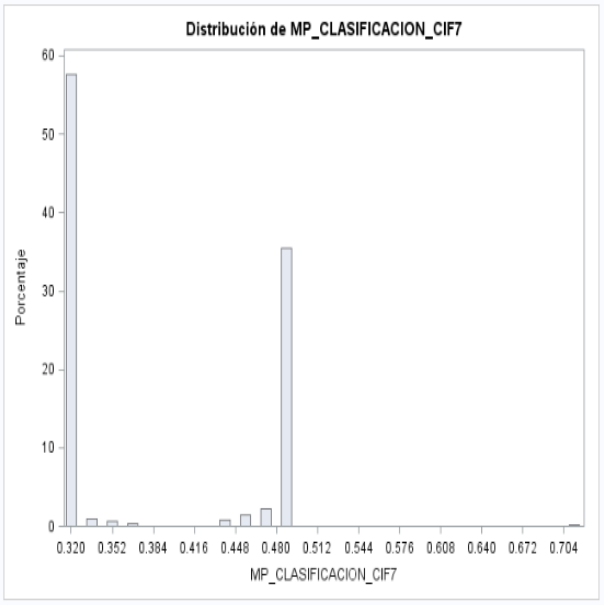

Distribución de Factor1
Porcentaje
Factor1
-0.56 -0.32 -0.08 0.16 0.40 0.64 0.88 1.12 1.36 1.60 1.84 2.08 2.32 2.56 2.80

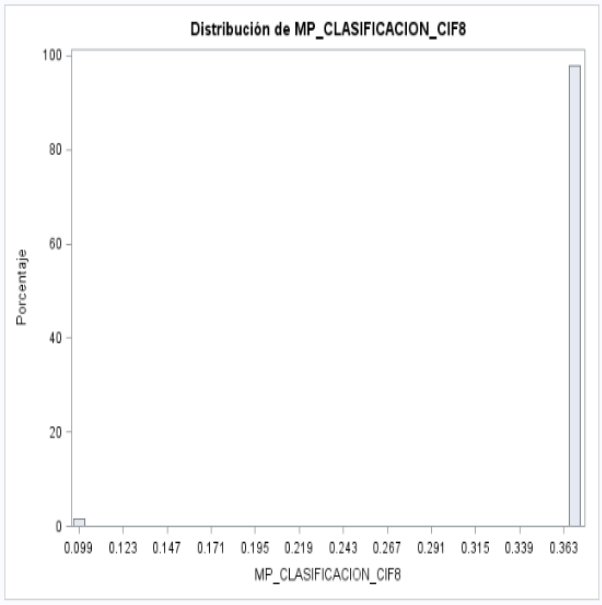

Distribución de Factor1
Porcentaje
Factor1
-0.56 -0.32 -0.08 0.16 0.40 0.64 0.88 1.12 1.36 1.60 1.84 2.08 2.32 2.56 2.80

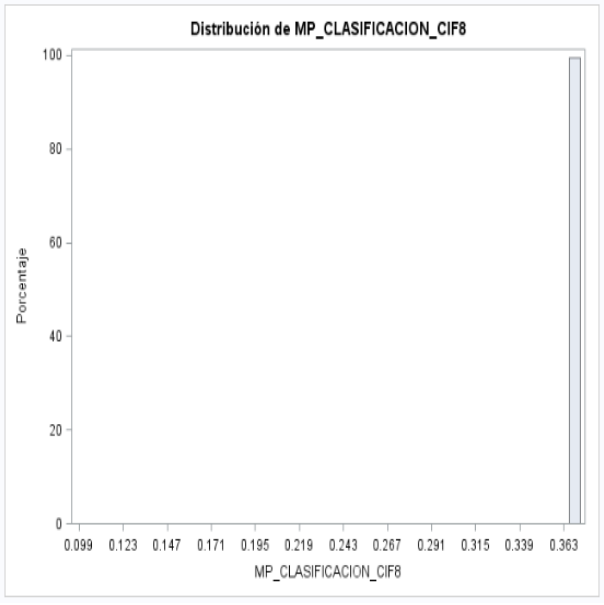

Distribución de MP_CLASIFICACION_CIF3
Porcentaje
MP_CLASIFICACION_CIF3
0.084 0.132 0.180 0.228 0.276 0.324 0.372 0.420 0.468 0.516 0.564 0.612

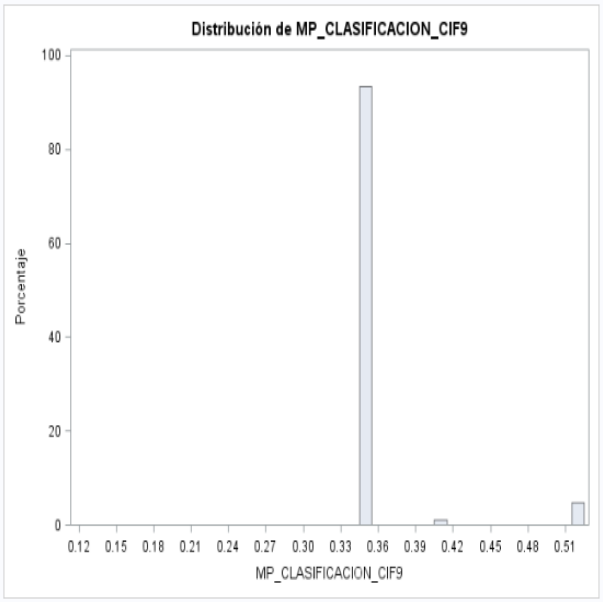

Distribución de MP_CLASIFICACION_CIF3
Porcentaje
MP_CLASIFICACION_CIF3
0.280 0.312 0.344 0.376 0.408 0.440 0.472 0.504 0.536 0.568 0.600 0.632

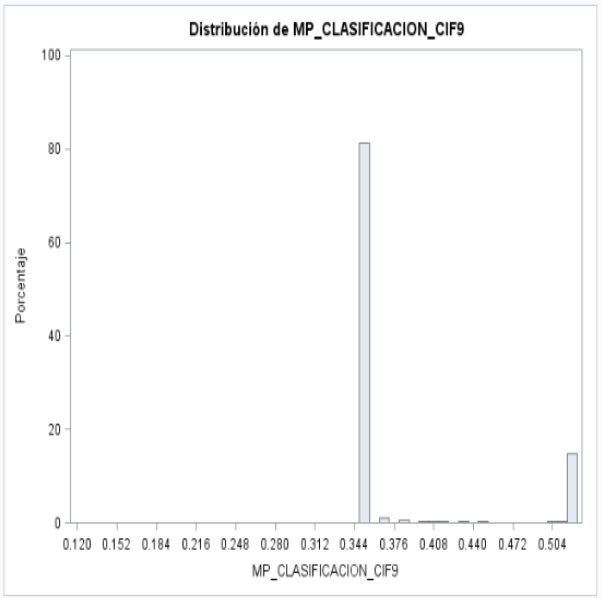

Distribución de MP_CLASIFICACION_CIF7
Porcentaje
0
20
40
60
80
100
0.320 0.352 0.384 0.416 0.448 0.480 0.512 0.544 0.576 0.608 0.640 0.672 0.704
MP_CLASIFICACION_CIF7

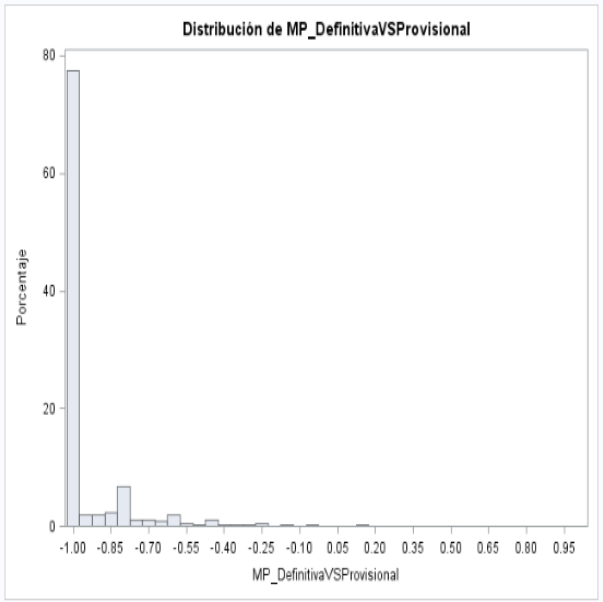

Distribución de MP_CLASIFICACION_CIF7
Porcentaje
0
10
20
30
40
50
60
0.320 0.352 0.384 0.416 0.448 0.480 0.512 0.544 0.576 0.608 0.640 0.672 0.704
MP_CLASIFICACION_CIF7

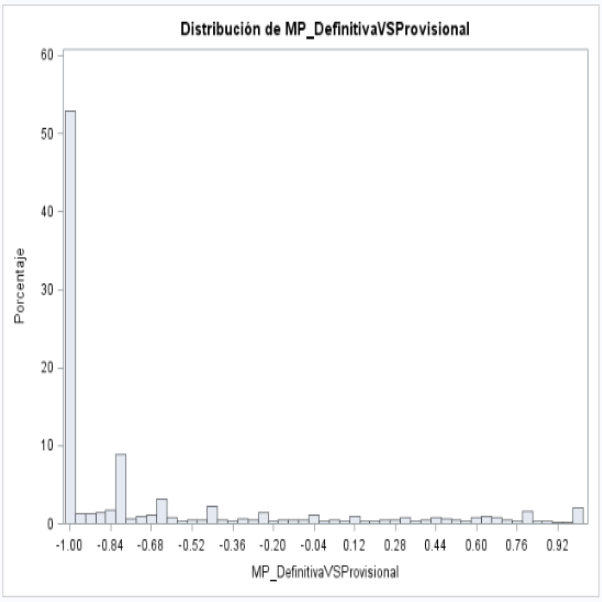

Distribución de MP_CLASIFICACION_CIF8
Porcentaje
0
20
40
60
80
100
0.099 0.123 0.147 0.171 0.195 0.219 0.243 0.267 0.291 0.315 0.339 0.363
MP_CLASIFICACION_CIF8

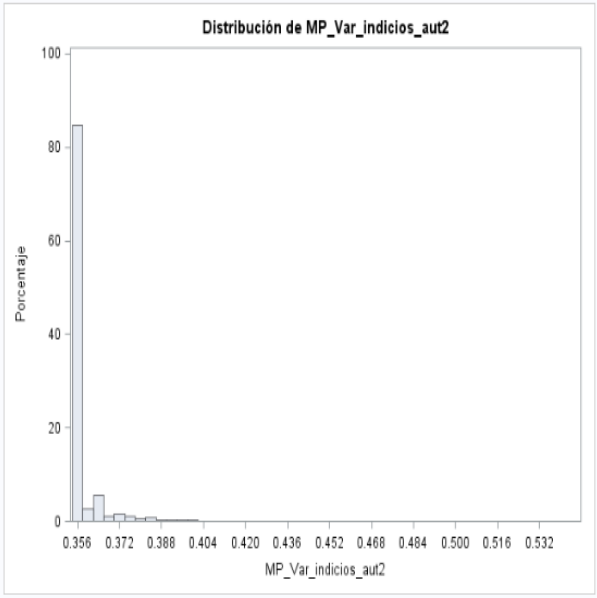

Distribución de MP_CLASIFICACION_CIF8
Porcentaje
0
20
40
60
80
100
0.099 0.123 0.147 0.171 0.195 0.219 0.243 0.267 0.291 0.315 0.339 0.363
MP_CLASIFICACION_CIF8

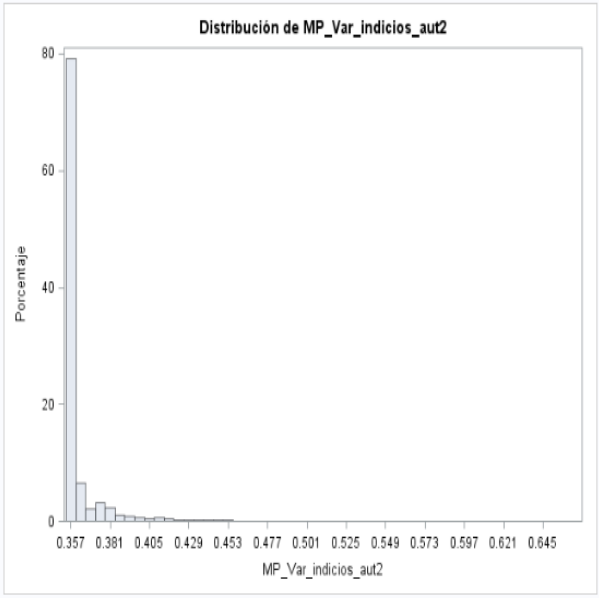

Distribución de MP_CLASIFICACION_CIF9
Porcentaje
MP_CLASIFICACION_CIF9

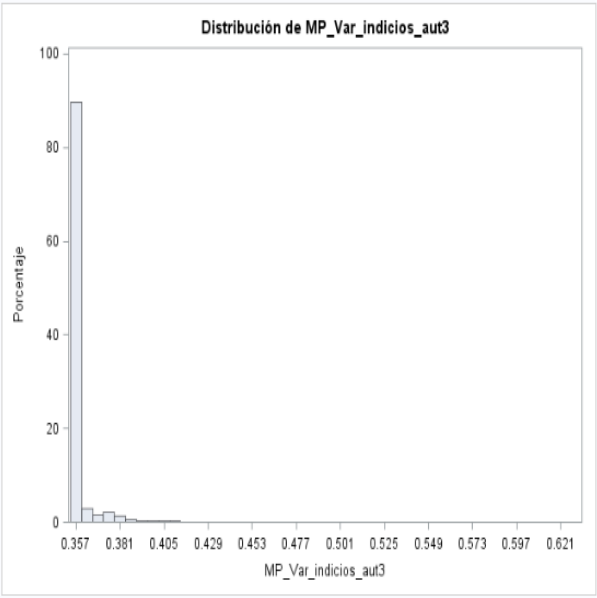

Distribución de MP_CLASIFICACION_CIF9
Porcentaje
MP_CLASIFICACION_CIF9

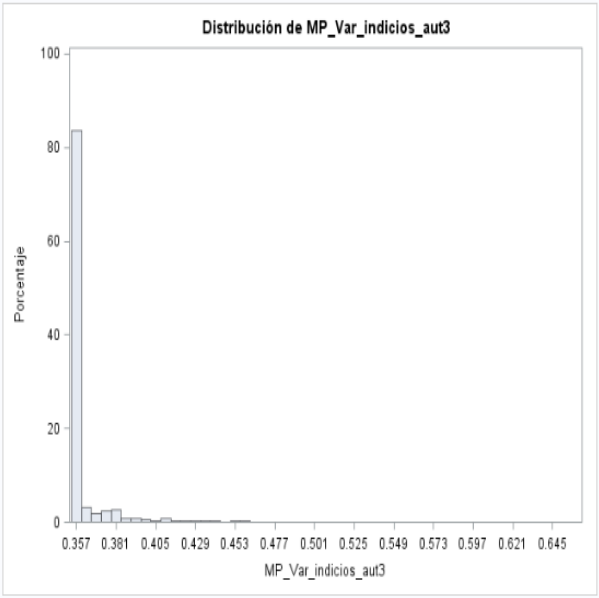

Distribución de MP_DefinitivaVSProvisional
Porcentaje
MP_DefinitivaVSProvisional

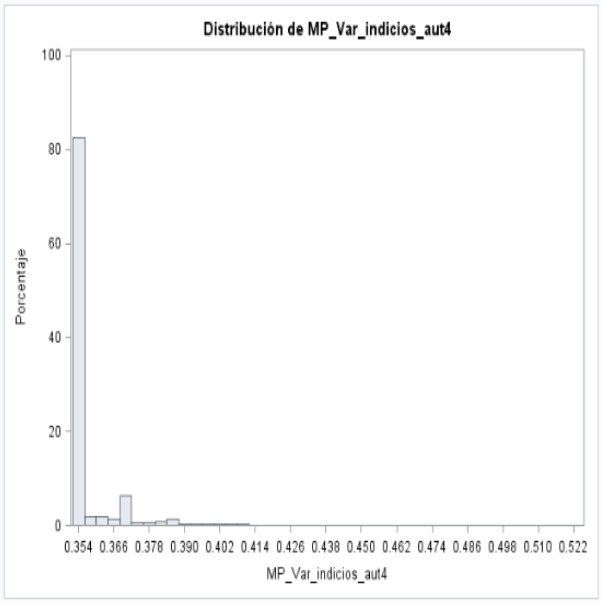

Distribución de MP_DefinitivaVSProvisional
Porcentaje
MP_DefinitivaVSProvisional

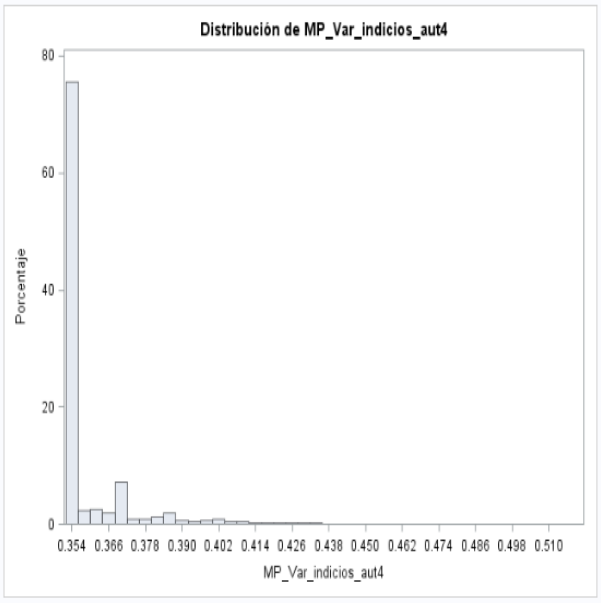

Distribución de MP_Var_indicios_aut2
Porcentaje
0
20
40
60
80
100
0.356 0.372 0.388 0.404 0.420 0.436 0.452 0.468 0.484 0.500 0.516 0.532
MP_Var_indicios_aut2

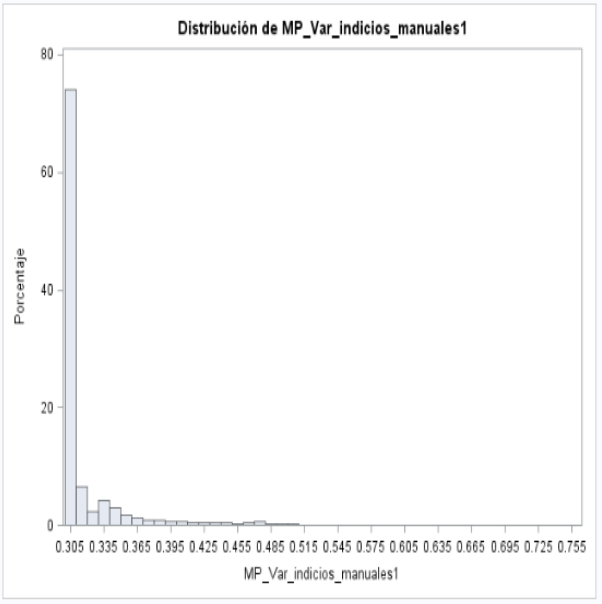

Distribución de MP_Var_indicios_aut2
Porcentaje
0
20
40
60
80
0.357 0.381 0.405 0.429 0.453 0.477 0.501 0.525 0.549 0.573 0.597 0.621 0.645
MP_Var_indicios_aut2

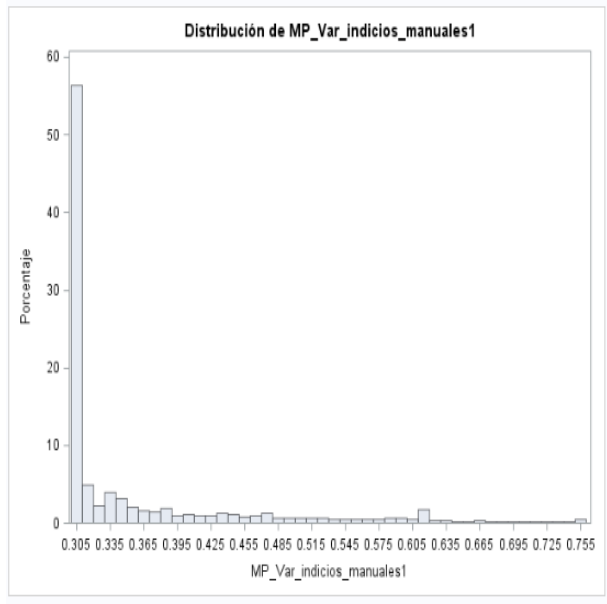

Distribución de MP_Var_indicios_aut3
Porcentaje
0
20
40
60
80
100
0.357 0.381 0.405 0.429 0.453 0.477 0.501 0.525 0.549 0.573 0.597 0.621
MP_Var_indicios_aut3

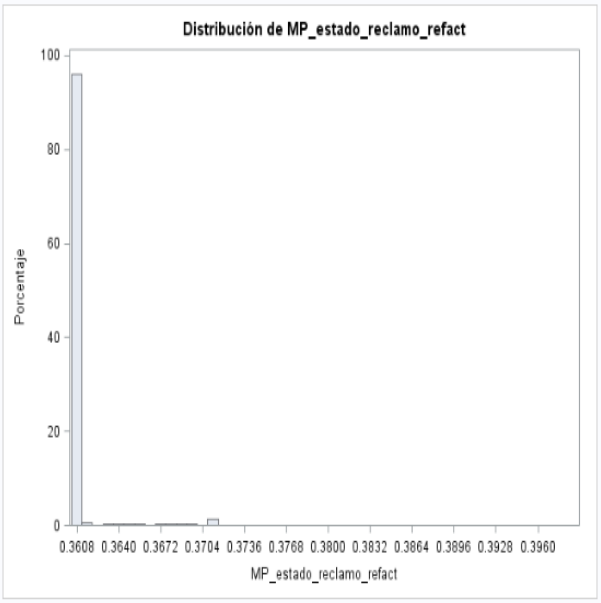

Distribución de MP_Var_indicios_aut3
Porcentaje
0
20
40
60
80
100
0.357 0.381 0.405 0.429 0.453 0.477 0.501 0.525 0.549 0.573 0.597 0.621 0.645
MP_Var_indicios_aut3

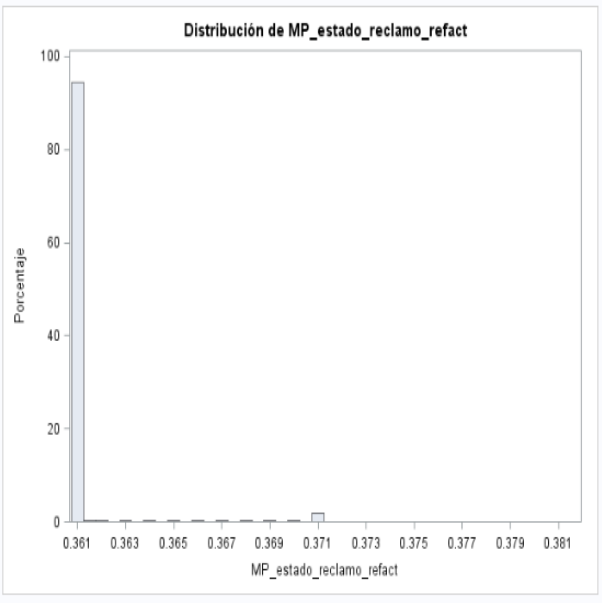

Distribución de MP_Var_indicios_aut4
Porcentaje
0 20 40 60 80 100
0.354 0.366 0.378 0.390 0.402 0.414 0.426 0.438 0.450 0.462 0.474 0.486 0.498 0.510 0.522
MP_Var_indicios_aut4

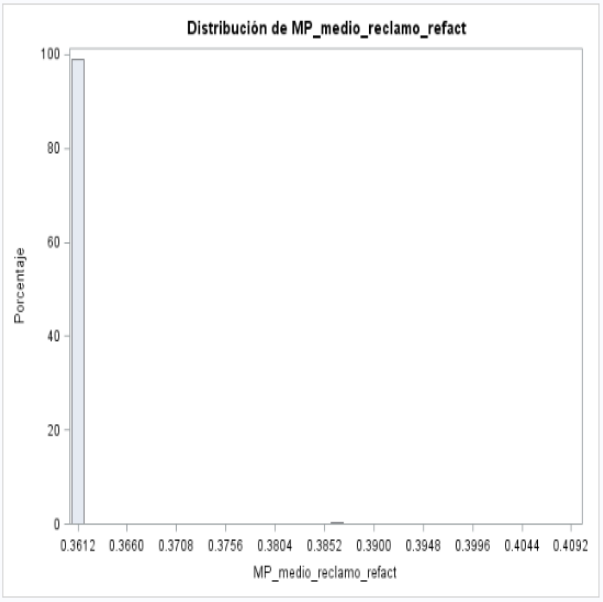

Distribución de MP_Var_indicios_aut4
Porcentaje
0 20 40 60 80
0.354 0.366 0.378 0.390 0.402 0.414 0.426 0.438 0.450 0.462 0.474 0.486 0.498 0.510
MP_Var_indicios_aut4

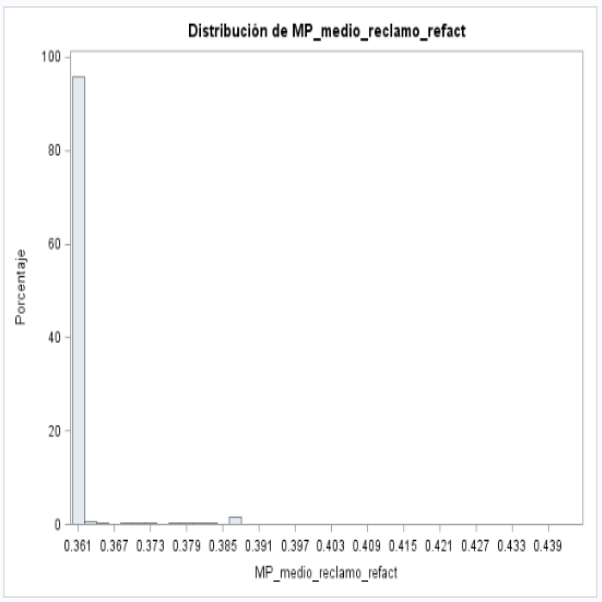

Distribución de MP_Var_indicios_manuales1
Porcentaje
0 20 40 60 80
0.305 0.335 0.365 0.395 0.425 0.455 0.485 0.515 0.545 0.575 0.605 0.635 0.665 0.695 0.725 0.755
MP_Var_indicios_manuales1

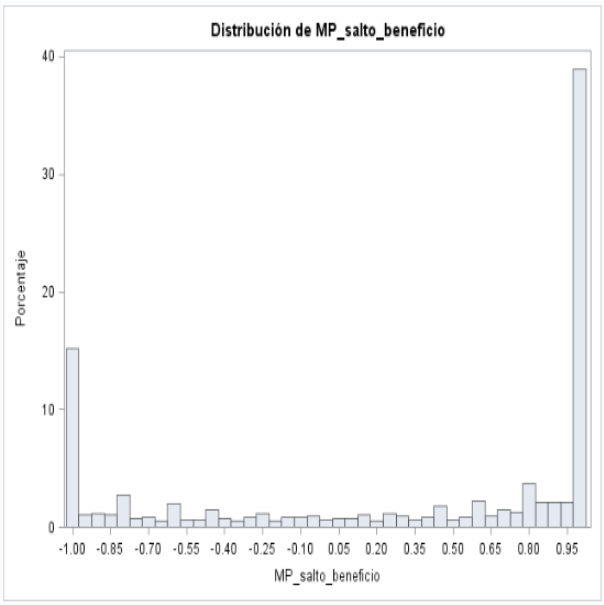

Distribución de MP_Var_indicios_manuales1
Porcentaje
0 10 20 30 40 50 60
0.305 0.335 0.365 0.395 0.425 0.455 0.485 0.515 0.545 0.575 0.605 0.635 0.665 0.695 0.725 0.755
MP_Var_indicios_manuales1

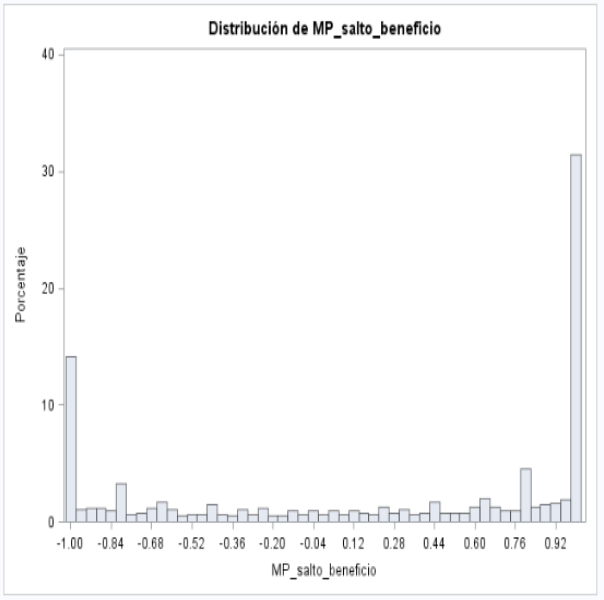

Distribución de MP_estado_reclamo_refact
Porcentaje
100
80
60
40
20
0
0.3608 0.3640 0.3672 0.3704 0.3736 0.3768 0.3800 0.3832 0.3864 0.3896 0.3928 0.3960
MP_estado_reclamo_refact

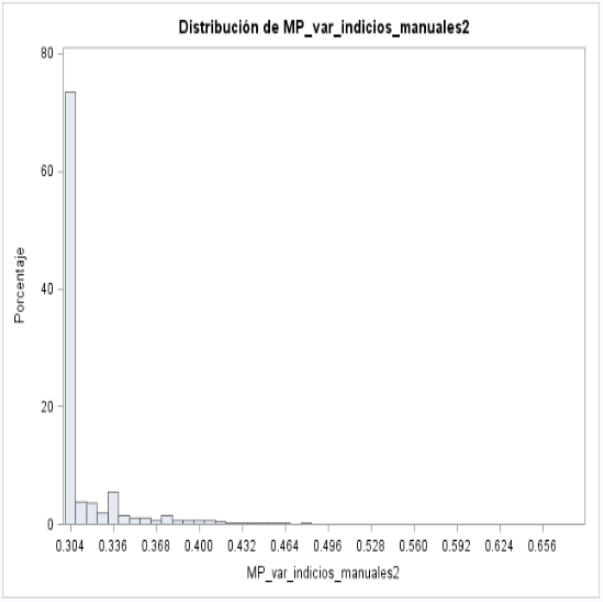

Distribución de MP_estado_reclamo_refact
Porcentaje
100
80
60
40
20
0
0.361 0.363 0.365 0.367 0.369 0.371 0.373 0.375 0.377 0.379 0.381
MP_estado_reclamo_refact

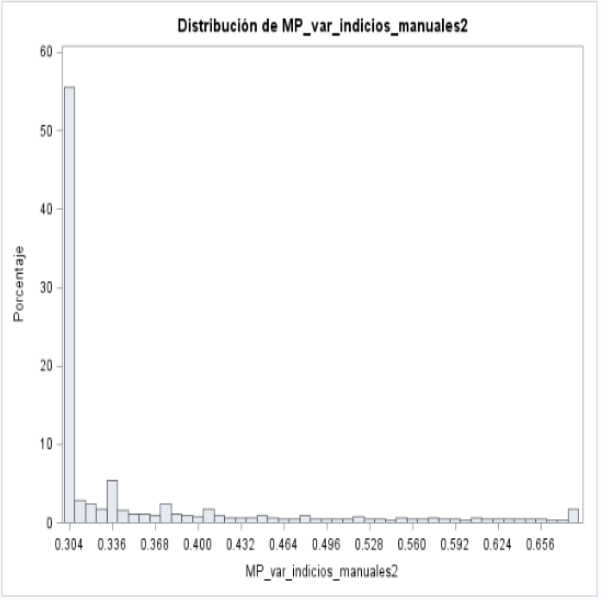

Distribución de MP_medio_reclamo_refact
Porcentaje
100
80
60
40
20
0
0.3612 0.3660 0.3708 0.3756 0.3804 0.3852 0.3900 0.3948 0.3996 0.4044 0.4092
MP_medio_reclamo_refact

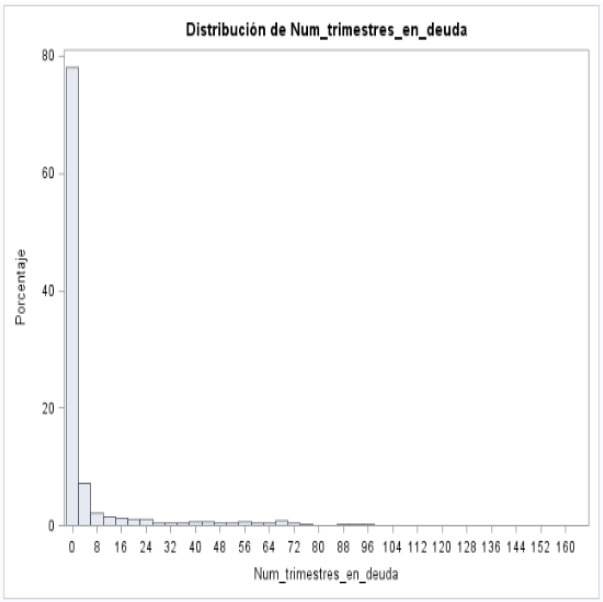

Distribución de MP_medio_reclamo_refact
Porcentaje
100
80
60
40
20
0
0.361 0.367 0.373 0.379 0.385 0.391 0.397 0.403 0.409 0.415 0.421 0.427 0.433 0.439
MP_medio_reclamo_refact

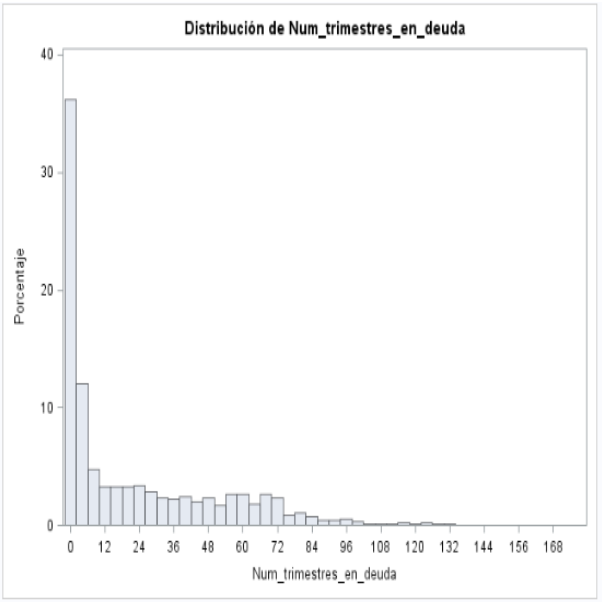

Distribución de MP_salto_beneficio
Porcentaje
0
10
20
30
40
-1.00 -0.85 -0.70 -0.55 -0.40 -0.25 -0.10 0.05 0.20 0.35 0.50 0.65 0.80 0.95
MP_salto_beneficio

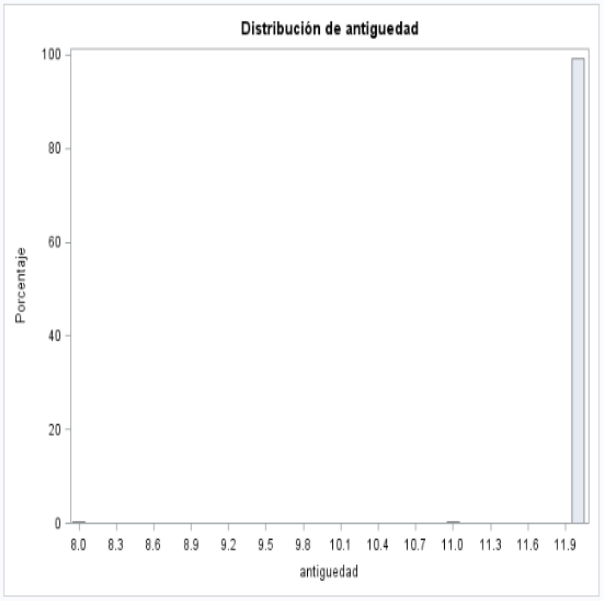

Distribución de MP_salto_beneficio
Porcentaje
0
10
20
30
40
-1.00 -0.84 -0.68 -0.52 -0.36 -0.20 -0.04 0.12 0.28 0.44 0.60 0.76 0.92
MP_salto_beneficio

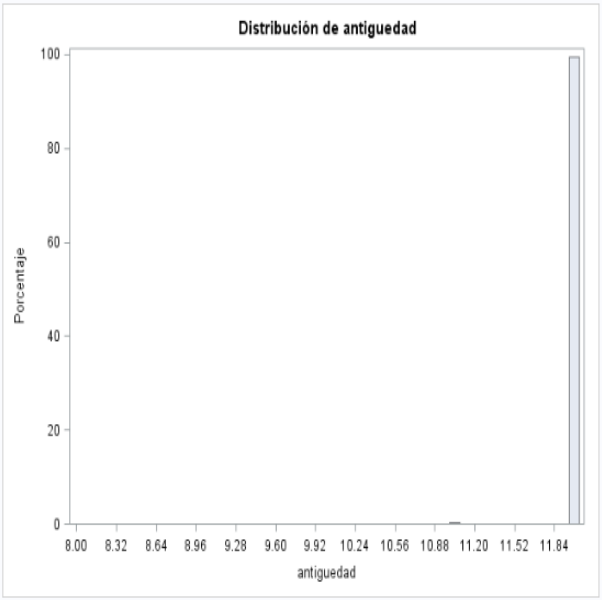

Distribución de MP_var_indicios_manuales2
Porcentaje
0
20
40
60
80
0.304 0.336 0.368 0.400 0.432 0.464 0.496 0.528 0.560 0.592 0.624 0.656
MP_var_indicios_manuales2

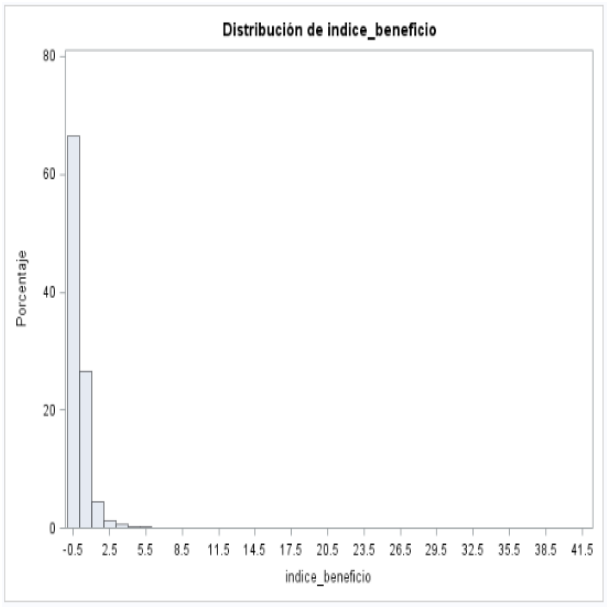

Distribución de MP_var_indicios_manuales2
Porcentaje
0
10
20
30
40
50
60
0.304 0.336 0.368 0.400 0.432 0.464 0.496 0.528 0.560 0.592 0.624 0.656
MP_var_indicios_manuales2

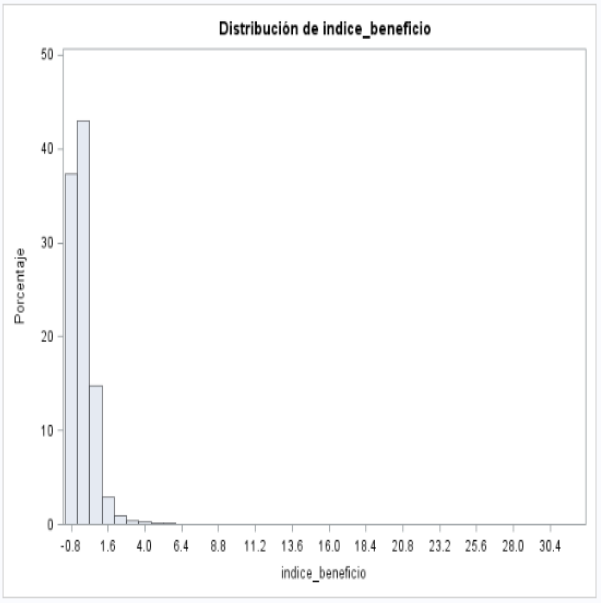

Distribución de Num_trimestres_en_deuda
Porcentaje
0
20
40
60
80
0 8 16 24 32 40 48 56 64 72 80 88 96 104 112 120 128 136 144 152 160
Num_trimestres_en_deuda

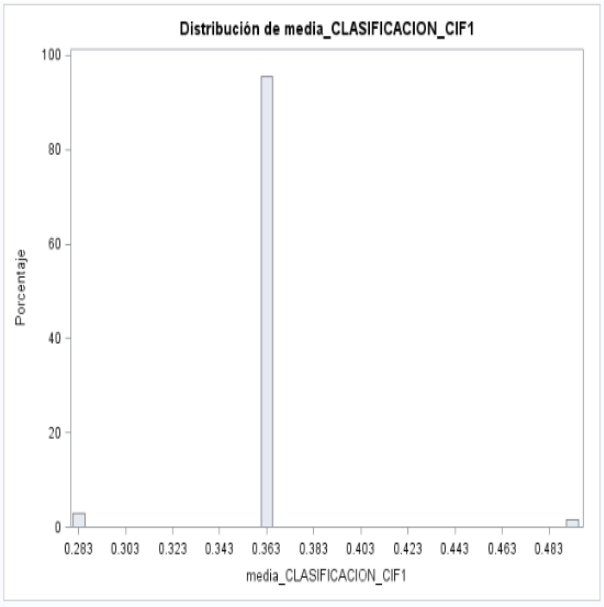

Distribución de Num_trimestres_en_deuda
Porcentaje
0
10
20
30
40
0 12 24 36 48 60 72 84 96 108 120 132 144 156 168
Num_trimestres_en_deuda

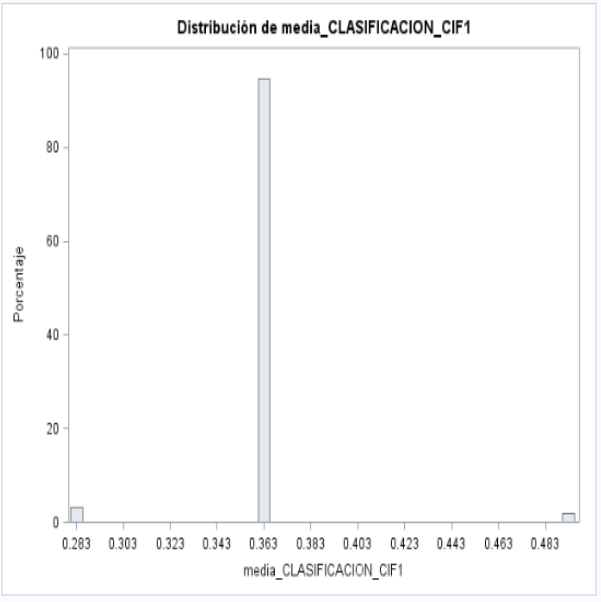

Distribución de antiguedad
Porcentaje
0
20
40
60
80
100
8.0 8.3 8.6 8.9 9.2 9.5 9.8 10.1 10.4 10.7 11.0 11.3 11.6 11.9
antiguedad

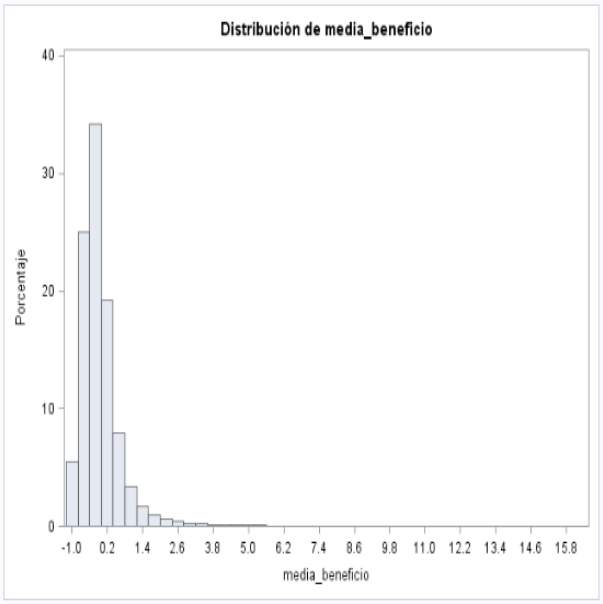

Distribución de antiguedad
Porcentaje
0
20
40
60
80
100
8.00 8.32 8.64 8.96 9.28 9.60 9.92 10.24 10.56 10.88 11.20 11.52 11.84
antiguedad

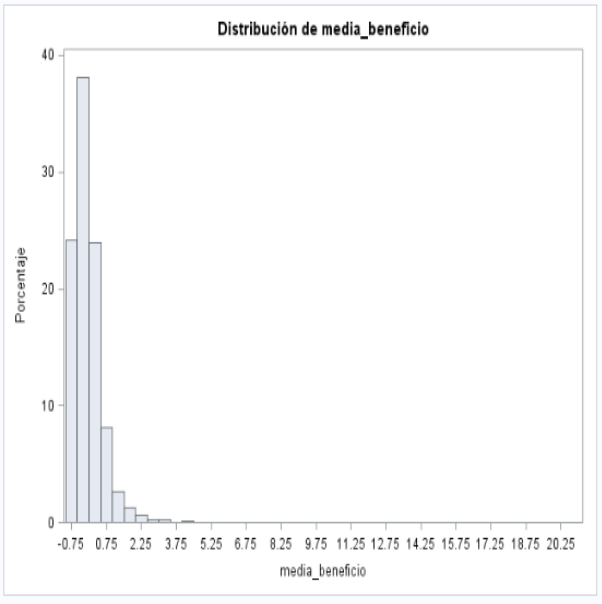

Distribución de indice_beneficio
Porcentaje
0
20
40
60
80
-0.5
2.5
5.5
8.5
11.5
14.5
17.5
20.5
23.5
26.5
29.5
32.5
35.5
38.5
41.5
indice_beneficio

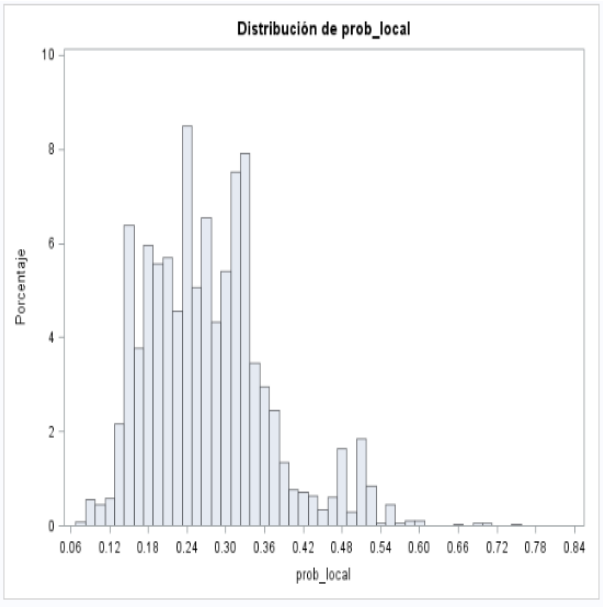

Distribución de indice_beneficio
Porcentaje
0
10
20
30
40
50
-0.8
1.6
4.0
6.4
8.8
11.2
13.6
16.0
18.4
20.8
23.2
25.6
28.0
30.4
indice_beneficio

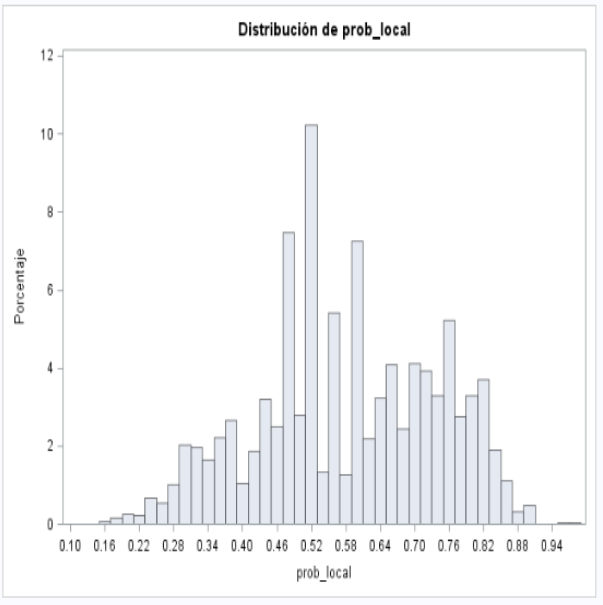

Distribución de media_CLASIFICACION_CIF1
Porcentaje
0
20
40
60
80
100
0.283
0.303
0.323
0.343
0.363
0.383
0.403
0.423
0.443
0.463
0.483
media_CLASIFICACION_CIF1

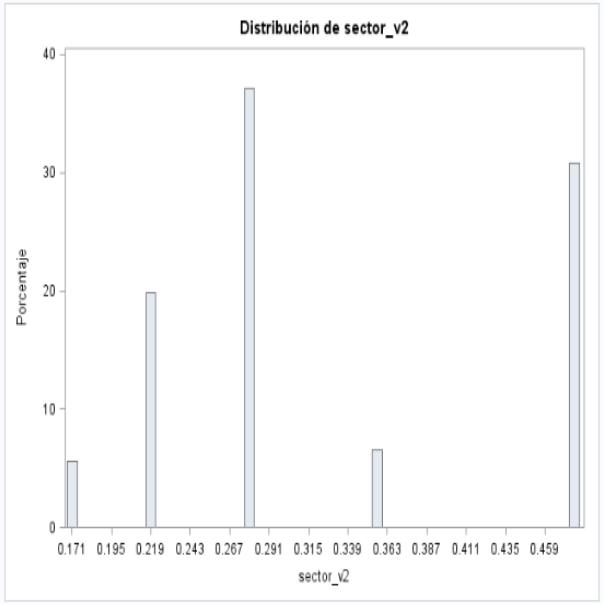

Distribución de media_CLASIFICACION_CIF1
Porcentaje
0
20
40
60
80
100
0.283
0.303
0.323
0.343
0.363
0.383
0.403
0.423
0.443
0.463
0.483
media_CLASIFICACION_CIF1

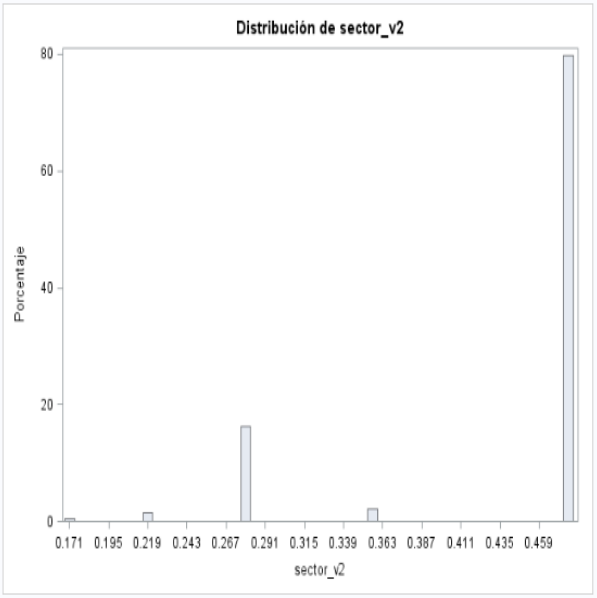

Distribución de media_beneficio
Porcentaje
media_beneficio
-1.0 0.2 1.4 2.6 3.8 5.0 6.2 7.4 8.6 9.8 11.0 12.2 13.4 14.6 15.8

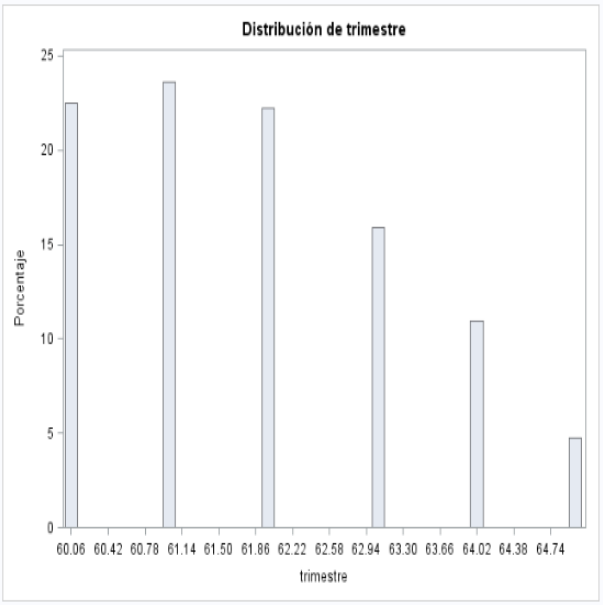

Distribución de media_beneficio
Porcentaje
media_beneficio
-0.75 0.75 2.25 3.75 5.25 6.75 8.25 9.75 11.25 12.75 14.25 15.75 17.25 18.75 20.25

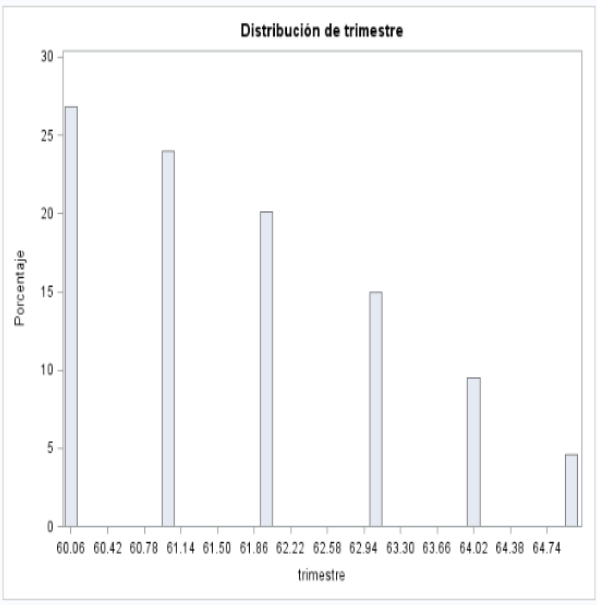

Distribución de prob_local
Porcentaje
prob_local
0.06 0.12 0.18 0.24 0.30 0.36 0.42 0.48 0.54 0.60 0.66 0.72 0.78 0.84

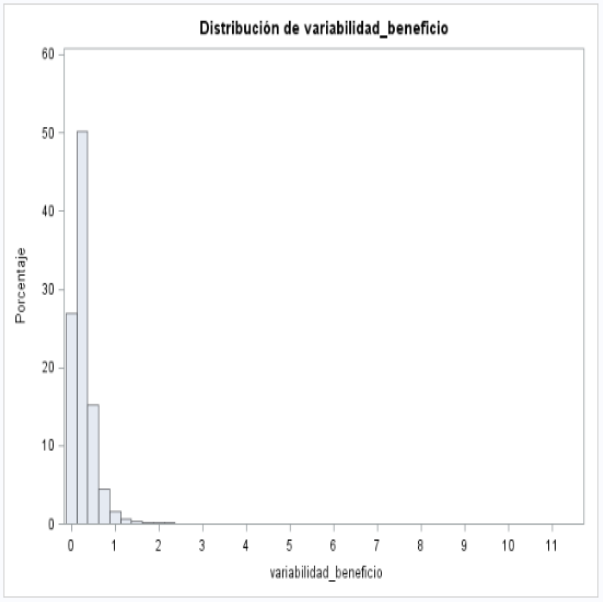

Distribución de prob_local
Porcentaje
prob_local
0.10 0.16 0.22 0.28 0.34 0.40 0.46 0.52 0.58 0.64 0.70 0.76 0.82 0.88 0.94

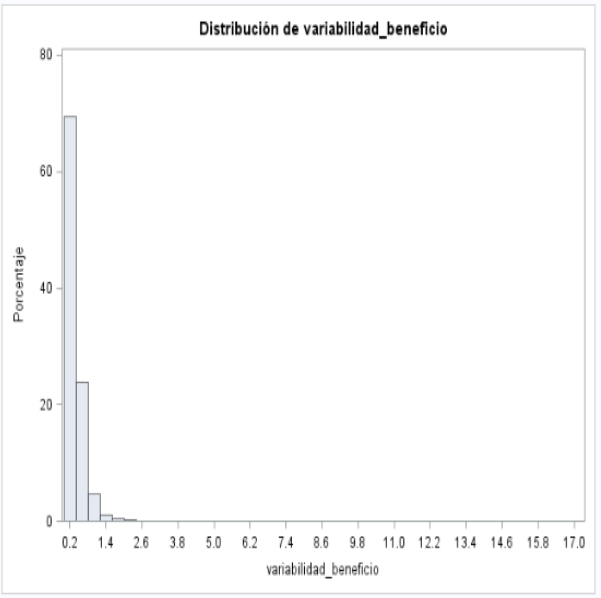

Distribución de sector_v2
Porcentaje
0
10
20
30
40
0.171 0.195 0.219 0.243 0.267 0.291 0.315 0.339 0.363 0.387 0.411 0.435 0.459
sector_v2


Distribución de sector_v2
Porcentaje
0
20
40
60
80
0.171 0.195 0.219 0.243 0.267 0.291 0.315 0.339 0.363 0.387 0.411 0.435 0.459
sector_v2


Distribución de trimestre
Porcentaje
0
5
10
15
20
25
60.06 60.42 60.78 61.14 61.50 61.86 62.22 62.58 62.94 63.30 63.66 64.02 64.38 64.74
trimestre


Distribución de trimestre
Porcentaje
0
5
10
15
20
25
30
60.06 60.42 60.78 61.14 61.50 61.86 62.22 62.58 62.94 63.30 63.66 64.02 64.38 64.74
trimestre





11.6. Código.

Se prescinde del código empleado para la formación de la Base de Datos dado que revelaría información de la empresa suministradora de los datos.

```
*------Logistica-----*/

libname muestra "C:\Users\Dani\Desktop\AaTFM\Datos TFM\Muestra";run;
libname datos "C:\Users\DepEIOIII\Desktop\dani\Totales\Totales";run;


proc printto print="C:\Users\DepEIOIII\Desktop\dani\out_logistica.txt";run;
proc printto log="C:\Users\DepEIOIII\Desktop\dani\log_logistica.txt";run;

data datos.resumen_logistica; format Metodo $10. Funcion $10.; informat Metodo Funcion
$10.; put Metodo $ Funcion $;run;


%macro logistica_STEPWISE;
%let lista='0.1 0.05 0.01 0.001 0.0001 0.00001 0.000001';
%let nume=7;
%do  i=1 %to &nume %by 1;
  data _null_;p_entrada=scanq(&lista,&i);call symput('p_entrada',left(p_entrada));run;
  %let lista2='Probit Logit Cloglog';
  %let nume2=3;
  %do k=1 %to &nume2 %by 1;
    data _null_;funcion=scanq(&lista2,&k);call symput('funcion',left(funcion));run;

    proc logistic data=datos.aprendizaje outest=parametros;
```

```
    class CLASIFICACION_CIF6_V2 TENER_ESTIMADO_FRAUDE_GRUPO CLASIFICACION_CIF5_V2
ESTM_FRAUDE_GRUP_TRIM_V2;
    model fraude (event='1')= CLASIFICACION_CIF6_v2 CLASIFICACION_CIF5_v2
Tener_estimado_fraude_grupo estm_fraude_grup_trim_v2
    F1 F2      F4      F6      F7      F8      Factor1 MP_CLASIFICACION_CIF3
       MP_CLASIFICACION_CIF7 MP_CLASIFICACION_CIF8
    MP_CLASIFICACION_CIF9     MP_DefinitivaVSProvisional    MP_Var_indicios_aut2
       MP_Var_indicios_aut3   MP_Var_indicios_aut4
    MP_Var_indicios_manuales1 MP_estado_reclamo_refact      MP_medio_reclamo_refact
       MP_salto_beneficio     MP_var_indicios_manuales2
    Num_trimestres_en_deuda   antiguedad     indice_beneficio
       media_CLASIFICACION_CIF1      media_beneficio       prob_local     sector_v2
    trimestre  variabilidad_beneficio

    / selection=FORWARD Link=&funcion SLSTAY=&p_entrada SLENTRY=&p_entrada;
    score data=datos.validacion out=salpredi;
    run;
    %do pmax=8 %to 10 %by 1;
         %do pmin=2 %to 7 %by 1;
        data tabla1 (keep=P 1 fraude); set salpredi; run;
        data tabla2 (drop=P 1); set tabla1; if ((&pmin)/10<P_1<(&pmax)/10) then
fraude_estimado=1; else fraude_estimado=-1; run;

        proc sql noprint; create table VP as select count(*) as VP from tabla2 where
(fraude=-1 and fraude_estimado=-1);quit;
        proc sql noprint; create table FP as select count(*) as FP from tabla2 where
(fraude=-1 and fraude_estimado=1);quit;
        proc sql noprint; create table FN as select count(*) as FN from tabla2 where
(fraude=1 and fraude_estimado=-1);quit;
        proc sql noprint; create table VN as select count(*) as VN from tabla2 where
(fraude=1 and fraude_estimado=1);quit;

               data parametros (drop=_LINK_ _TYPE_ _STATUS_ _NAME_ Intercept _LNLIKE_
_ESTTYPE_); set parametros; run;
        proc transpose data=parametros out=parametros tras;run;
        data parametros_tras (keep=COL1); set parametros_tras;if (COL1=.) then delete;
run;
        proc sql noprint; create table num_parametros as select count(*) as
num_parametros from parametros_tras;quit;
        data criterio; merge VP FP FN VN num_parametros; run;

        data criterio; set criterio;
        Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;

        data criterio;set criterio; format Metodo $10. Funcion $10.; informat Metodo
Funcion $10.; put Metodo $ Funcion $;run;
        data criterio;set criterio;
        Metodo='STEPWISE';
        Funcion="&funcion";
           p_entrada=&p_entrada;
           p_salida=&p_entrada;
        pmax=(&pmax)/10;
               pmin=(&pmin)/10;
        run;
               data criterio;set criterio; format Metodo $10. Funcion $10.; informat
Metodo Funcion $10.; put Metodo $ Funcion $;run;
        data datos.Resumen_logistica; set datos.Resumen_logistica criterio;run;
               %end;
         %end;
    %end;
%end;
%mend;

%logistica_STEPWISE;

%macro logistica_FORWARD;
%let lista='0.1 0.05 0.01 0.001 0.0001 0.00001 0.000001';
%let nume=7;
%do  i=1 %to &nume %by 1;
  data _null_;p_entrada=scanq(&lista,&i);call symput('p_entrada',left(p_entrada));run;
```

```
  %let lista2='Probit Logit Cloglog';
  %let nume2=3;
  %do j=1 %to &nume2 %by 1;
    data _null_;funcion=scanq(&lista2,&j);call symput('funcion',left(funcion));run;

    proc logistic data=datos.aprendizaje outest=parametros;
    class CLASIFICACION_CIF6_V2 TENER_ESTIMADO_FRAUDE_GRUPO CLASIFICACION_CIF5_V2
ESTM_FRAUDE_GRUP_TRIM_V2;
    model fraude (event='1')= CLASIFICACION_CIF6_v2 CLASIFICACION_CIF5_v2
Tener_estimado_fraude_grupo estm_fraude_grup_trim_v2
    F1  F2      F4      F6      F7      F8      Factor1 MP_CLASIFICACION_CIF3
       MP_CLASIFICACION_CIF7  MP_CLASIFICACION_CIF8
    MP_CLASIFICACION_CIF9     MP_DefinitivaVSProvisional    MP_Var_indicios_aut2
       MP_Var_indicios_aut3   MP_Var_indicios_aut4
    MP_Var_indicios_manuales1 MP_estado_reclamo_refact      MP_medio_reclamo_refact
       MP_salto_beneficio     MP_var_indicios_manuales2
    Num_trimestres_en_deuda   antiguedad     indice_beneficio
       media_CLASIFICACION_CIF1      media_beneficio       prob_local     sector_v2
    trimestre  variabilidad_beneficio

       / selection=FORWARD Link=&funcion SLENTRY=&p_entrada;
    score data=datos.validacion out=salpredi;
    run;
    %do pmax=8 %to 10 %by 1;
         %do pmin=2 %to 7 %by 1;
        data tabla1 (keep=P 1 fraude); set salpredi; run;
        data tabla2 (drop=P 1); set tabla1; if ((&pmin)/10<P 1<(&pmax)/10) then
fraude_estimado=1; else fraude_estimado=-1; run;/*-P_1-> probabilidad de fraude
estimada-*/

        proc sql noprint; create table VP as select count(*) as VP from tabla2 where
(fraude=-1 and fraude_estimado=-1);quit;
        proc sql noprint; create table FP as select count(*) as FP from tabla2 where
(fraude=-1 and fraude_estimado=1);quit;
        proc sql noprint; create table FN as select count(*) as FN from tabla2 where
(fraude=1 and fraude_estimado=-1);quit;
        proc sql noprint; create table VN as select count(*) as VN from tabla2 where
(fraude=1 and fraude_estimado=1);quit;

               data parametros (drop=_LINK_ _TYPE_ _STATUS_ _NAME_ Intercept _LNLIKE_
_ESTTYPE_); set parametros; run;
        proc transpose data=parametros out=parametros tras;run;
        data parametros_tras (keep=COL1); set parametros_tras;if (COL1=.) then delete;
run;
        proc sql noprint; create table num_parametros as select count(*) as
num_parametros from parametros_tras;quit;
        data criterio; merge VP FP FN VN num_parametros; run;

        data criterio; set criterio;
        Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;

        data criterio;set criterio; format Metodo $10. Funcion $10.; informat Metodo
Funcion $10.; put Metodo $ Funcion $;run;
        data criterio;set criterio;
        Metodo='FORWARD';
        Funcion="&funcion";
           p entrada=&p_entrada;
           p_salida=0;
        pmax=(&pmax)/10;
               pmin=(&pmin)/10;
        run;
               data criterio;set criterio; format Metodo $10. Funcion $10.; informat
Metodo Funcion $10.; put Metodo $ Funcion $;run;
        data datos.Resumen_logistica; set datos.Resumen_logistica criterio;run;
         %end;
    %end;
  %end;
%end;
%mend;
```

```
%logistica_FORWARD;


%macro logistica_BACKWARD;
%let lista='0.1 0.05 0.01 0.001 0.0001 0.00001 0.000001';
%let nume=7;
%do  i=1 %to &nume %by 1;
  data _null_;p_salida=scanq(&lista,&i);call symput('p_salida',left(p_salida));run;
  %let lista2='Probit Logit Cloglog';
  %let nume2=3;
  %do j=1 %to &nume2 %by 1;
    data _null_;funcion=scanq(&lista2,&j);call symput('funcion',left(funcion));run;

     proc logistic data=datos.aprendizaje outest=parametros;
     class CLASIFICACION_CIF6_V2 TENER_ESTIMADO_FRAUDE_GRUPO CLASIFICACION_CIF5_V2
ESTM_FRAUDE_GRUP_TRIM_V2;
     model fraude (event='1')= CLASIFICACION_CIF6_v2 CLASIFICACION_CIF5_v2
Tener_estimado_fraude_grupo estm_fraude_grup_trim_v2
     F1 F2      F4      F6      F7      F8      Factor1 MP_CLASIFICACION_CIF3
       MP_CLASIFICACION_CIF7 MP_CLASIFICACION_CIF8
     MP_CLASIFICACION_CIF9    MP_DefinitivaVSProvisional    MP_Var_indicios_aut2
       MP_Var_indicios_aut3  MP_Var_indicios_aut4
     MP_Var_indicios_manuales1       MP_estado_reclamo_refact
       MP_medio_reclamo_refact       MP_salto_beneficio    MP_var_indicios_manuales2
     Num_trimestres_en_deuda  antiguedad     indice_beneficio
       media_CLASIFICACION_CIF1      media_beneficio       prob_local     sector_v2
     trimestre variabilidad_beneficio CLASIFICACION_CIF6_V2*MP_CLASIFICACION_CIF3
CLASIFICACION_CIF5_V2*MP_CLASIFICACION_CIF3

        / selection=FORWARD Link=&funcion SLSTAY=&p_salida;
     score data=datos.validacion out=salpredi;
     run;
        %do pmax=8 %to 10 %by 1;
         %do pmin=2 %to 7 %by 1;
        data tabla1 (keep=P 1 fraude); set salpredi; run;
        data tabla2 (drop=P_1); set tabla1; if ((&pmin)/10<P_1<(&pmax)/10) then
fraude_estimado=1; else fraude_estimado=-1; run;

        proc sql noprint; create table VP as select count(*) as VP from tabla2 where
(fraude=-1 and fraude_estimado=-1);quit;
        proc sql noprint; create table FP as select count(*) as FP from tabla2 where
(fraude=-1 and fraude_estimado=1);quit;
        proc sql noprint; create table FN as select count(*) as FN from tabla2 where
(fraude=1 and fraude_estimado=-1);quit;
        proc sql noprint; create table VN as select count(*) as VN from tabla2 where
(fraude=1 and fraude_estimado=1);quit;

               data parametros (drop=_LINK_ _TYPE_ _STATUS_ _NAME_ Intercept _LNLIKE_
_ESTTYPE_); set parametros; run;
        proc transpose data=parametros out=parametros tras;run;
        data parametros_tras (keep=COL1); set parametros_tras;if (COL1=.) then delete;
run;
        proc sql noprint; create table num_parametros as select count(*) as
num_parametros from parametros_tras;quit;
        data criterio; merge VP FP FN VN num_parametros; run;

        data criterio; set criterio;
        Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;

        data criterio;set criterio; format Metodo $10. Funcion $10.; informat Metodo
Funcion $10.; put Metodo $ Funcion $;run;
        data criterio;set criterio;
        Metodo='BACKWARD';
        Funcion="&funcion";
           p_entrada=0;
           p salida=&p salida;
        pmax=(&pmax)/10;
               pmin=(&pmin)/10;
        run;
```

```
              data criterio;set criterio; format Metodo $10. Funcion $10.; informat
Metodo Funcion $10.; put Metodo $ Funcion $;run;
        data datos.Resumen_logistica; set datos.Resumen_logistica criterio;run;
         %end;
    %end;
  %end;
%end;
%mend;

%logistica_BACKWARD;

data datos.Resumen_logistica;set datos.Resumen_logistica;if _n_=1 then delete;run;

data datos.resumen_logistica; set datos.resumen_logistica;
por_aciertos=VN/(VN+FP);
por_visitas=(VN+FP)/(VN+FP+VP+FN);
n_visitas=por_visitas*320000;
run;



/*-------------------------------------------------------------*/
libname sal 'C:\Users\Dani\Desktop\AaTFM\Datos TFM';run;

proc logistic data=datos.aprendizaje outest=parametros;
class CLASIFICACION_CIF6_V2 TENER_ESTIMADO_FRAUDE_GRUPO CLASIFICACION_CIF5_V2
ESTM_FRAUDE_GRUP_TRIM_V2;
model fraude (event='1')= CLASIFICACION_CIF6_v2 CLASIFICACION_CIF5_v2
Tener_estimado_fraude_grupo estm_fraude_grup_trim_v2
F1      F2      F4      F6      F7      F8      Factor1 MP_CLASIFICACION_CIF3
        MP_CLASIFICACION_CIF7 MP_CLASIFICACION_CIF8
MP_CLASIFICACION_CIF9 MP_DefinitivaVSProvisional    MP_Var_indicios_aut2
        MP_Var_indicios_aut3  MP_Var_indicios_aut4
MP_Var_indicios_manuales1     MP_estado_reclamo_refact      MP_medio_reclamo_refact
        MP_salto_beneficio    MP_var_indicios_manuales2
Num_trimestres_en_deuda       antiguedad     indice_beneficio
        media_CLASIFICACION_CIF1      media_beneficio       prob_local     sector_v2
trimestre      variabilidad_beneficio CLASIFICACION_CIF6_V2*MP_CLASIFICACION_CIF3
CLASIFICACION_CIF5_V2*MP_CLASIFICACION_CIF3
/ selection=FORWARD Link=Cloglog SLENTRY=0.01;
score data=datos.validacion out=sal.salpredi;
run;

proc sort data=sal.salpredi; by fraude; run;

proc univariate data=sal.salpredi;
      var P_1;
      histogram;
         by fraude;
   run;



proc printto; run;
proc printto print=print;run;
proc sql;  select *  from datos.resumen_logistica where Matthews=(select max(Matthews)
from datos.resumen_logistica);quit;

proc sort data=datos.resumen_logistica ; by descending Matthews ;run;



proc sort data=datos.resumen_logistica ; by n_visitas ;run;


/*---Redes Neuronales---*/

libname muestra "C:\Users\DepEIOIII\Desktop\dani\Muestra"; run;
libname datos "C:\Users\DepEIOIII\Desktop\dani\Totales"; run;
libname resumen "H:\Master\TFM\Resumenes";run;
```

```
/*Generar Muestras*/

data muestra.aprendizaje; set datos.aprendizaje; aleatorio=rand("uniform");run; proc
sort data=muestra.aprendizaje; by aleatorio;run; data muestra.aprendizaje
(drop=aleatorio); set muestra.aprendizaje; if (_N_>6590) then delete; run;
data muestra.validacion; set datos.validacion; aleatorio=rand("uniform"); run;proc sort
data=muestra.validacion; by aleatorio;run; data muestra.validacion (drop=aleatorio); set
muestra.validacion; if (_N_>2210) then delete; run;
data muestra.test; set datos.test; aleatorio=rand("uniform"); run;proc sort
data=muestra.test; by aleatorio;run; data muestra.test (drop=aleatorio); set
muestra.test; if (_N_>640) then delete; run;

/*-Redes Neuronales-*/

/*-options mprint o options mprint=0^ o options nonotes mprint=0 ; te extiende y te
explica mejor el log*/
/* proc printto print="C:/ basuta.txt";run;
   proc printto; run;

proc catalog cat=gseg;run;----> elimina graficos*/

proc printto print="E:\Master\TFM\out_redes.txt";run;
proc printto log="E:\Master\TFM\log_redes.txt";run;
proc printto print=print;run;

PROC DMDB DATA=datos.aprendizaje dmdbcat=datos.catalogo_aprendizaje;
target fraude; var     F1      F2      F4      F6      F7      F8      Factor1
       MP_CLASIFICACION_CIF3 MP_CLASIFICACION_CIF7 MP_CLASIFICACION_CIF8
       MP_CLASIFICACION_CIF9 MP_DefinitivaVSProvisional    MP_Var_indicios_aut2
       MP_Var_indicios_aut3  MP_Var_indicios_aut4  MP_Var_indicios_manuales1
       MP_estado_reclamo_refact      MP_medio_reclamo_refact       MP_salto_beneficio
       MP_var_indicios_manuales2     Num_trimestres_en_deuda       antiguedad
       indice_beneficio      media_CLASIFICACION_CIF1      media_beneficio
       prob_local     sector_v2      trimestre      variabilidad_beneficio;
class fraude   CLASIFICACION_CIF6_v2 CLASIFICACION_CIF5_v2 Tener_estimado_fraude_grupo
estm_fraude_grup_trim_v2;run;

PROC DMDB DATA=muestra.aprendizaje dmdbcat=muestra.catalogo_aprendizaje;
target fraude; var     F1      F2      F4      F6      F7      F8      Factor1
       MP_CLASIFICACION_CIF3 MP_CLASIFICACION_CIF7 MP_CLASIFICACION_CIF8
       MP_CLASIFICACION_CIF9 MP_DefinitivaVSProvisional    MP_Var_indicios_aut2
       MP_Var_indicios_aut3  MP_Var_indicios_aut4  MP_Var_indicios_manuales1
       MP_estado_reclamo_refact      MP_medio_reclamo_refact       MP_salto_beneficio
       MP_var_indicios_manuales2     Num_trimestres_en_deuda       antiguedad
       indice_beneficio      media_CLASIFICACION_CIF1      media_beneficio
       prob_local     sector_v2      trimestre      variabilidad_beneficio;
class fraude   CLASIFICACION_CIF6_v2 CLASIFICACION_CIF5_v2 Tener_estimado_fraude_grupo
estm_fraude_grup_trim_v2;run;

data muestra.Resumen_train_1p; set muestra.Resumen_train_1p; if (nodos2=1) then delete;
run;

%macro redes;
data datos.resumen_train; format f_activacion $3.; informat f_activacion $3.; put
f_activacion;run;
%let lista='TAN EXP ARC ELL';
%let nume=4;
%do  i=1 %to &nume %by 1;
data  null ;activa=scanq(&lista,&i);call symput('activa',left(activa));run;
%do nodos2=1 %to 20 %by 1;
  %do nodos1=1 %to 20 %by 1;
        %do corte=3 %to 7 %by 1;
       proc neural data=muestra.aprendizaje dmdbcat=muestra.catalogo_aprendizaje;
       input F1      F2      F4      F6      F7      F8      Factor1 MP_CLASIFICACION_CIF3
       MP_CLASIFICACION_CIF7 MP_CLASIFICACION_CIF8 MP_CLASIFICACION_CIF9
       MP_DefinitivaVSProvisional    MP_Var_indicios_aut2  MP_Var_indicios_aut3
       MP_Var_indicios_aut4  MP_Var_indicios_manuales1     MP_estado_reclamo_refact
       MP_medio_reclamo_refact       MP_salto_beneficio     MP_var_indicios_manuales2
       Num_trimestres_en_deuda       antiguedad     indice_beneficio
```

```
       media_CLASIFICACION_CIF1      media_beneficio       prob_local     sector_v2
       trimestre      variabilidad_beneficio/ level=int;
        ;
        input CLASIFICACION_CIF6_v2 CLASIFICACION_CIF5_v2 Tener_estimado_fraude_grupo
estm_fraude_grup_trim_v2 / level=nom
        ;
        target  fraude;
        hidden &nodos1 / act=&activa id=H;
           hidden &nodos2 / act=&activa id=K;

        train tech=levmar;
        score data=muestra.aprendizaje out=salpredi outfit=salfit;
        run;

        data tabla1 (keep=P_fraude_1 fraude); set salpredi; run;
        data tabla2 (drop=P_fraude_1); set tabla1; if (P_fraude_1<&corte) then
fraude_estimado=1; else fraude_estimado=-1; run;

              proc sql noprint; create table VP as select count(*) as VP from tabla2
where (fraude=-1 and fraude_estimado=-1);quit;
              proc sql noprint; create table FP as select count(*) as FP from tabla2
where (fraude=-1 and fraude_estimado=1);quit;
              proc sql noprint; create table FN as select count(*) as FN from tabla2
where (fraude=1 and fraude_estimado=-1);quit;
              proc sql noprint; create table VN as select count(*) as VN from tabla2
where (fraude=1 and fraude_estimado=1);quit;

              data criterio; merge VP FP FN VN; run;
              data criterio; set criterio;
              Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));
              nodos1=&nodos1;
              nodos2=&nodos2;
              f_activacion='&activa';
              p_corte=&corte;
              run;

              data muestra.resumen_train_1p; set muestra.resumen_train_1p criterio;run;
     /* %end;*/
    %end;
  %end;
%end;
data muestra.resumen_train;muestra.resumen_train;if _n_=1 then delete;run;
proc print data=muestra.resumen_train;run;
%mend;

%redes;


/*------RED OPTIMA----*/

libname resumen "C:\Users\Dani\Dropbox\Master\TFM\Resumenes\Redes";run;
PROC DMDB DATA=datos.aprendizaje dmdbcat=datos.catalogo_aprendizaje;
target fraude; var     F1      F2      F4      F6      F7      F8      Factor1
       MP_CLASIFICACION_CIF3 MP_CLASIFICACION_CIF7 MP_CLASIFICACION_CIF8
       MP_CLASIFICACION_CIF9 MP_DefinitivaVSProvisional    MP_Var_indicios_aut2
       MP_Var_indicios_aut3  MP_Var_indicios_aut4  MP_Var_indicios_manuales1
       MP_estado_reclamo_refact      MP_medio_reclamo_refact       MP_salto_beneficio
       MP_var_indicios_manuales2     Num_trimestres_en_deuda       antiguedad
       indice_beneficio      media_CLASIFICACION_CIF1      media_beneficio
       prob_local     sector_v2      trimestre      variabilidad_beneficio;
class fraude   CLASIFICACION CIF6_v2 CLASIFICACION_CIF5_v2 Tener_estimado_fraude_grupo
estm_fraude_grup_trim_v2;run;

proc neural data=datos.aprendizaje dmdbcat=datos.catalogo_aprendizaje;
input F1       F2      F4      F6      F7      F8      Factor1 MP_CLASIFICACION_CIF3
       MP_CLASIFICACION_CIF7 MP_CLASIFICACION_CIF8 MP_CLASIFICACION_CIF9
       MP_DefinitivaVSProvisional    MP_Var_indicios_aut2  MP_Var_indicios_aut3
       MP_Var_indicios_aut4  MP_Var_indicios_manuales1     MP_estado_reclamo_refact
       MP_medio_reclamo_refact       MP_salto_beneficio     MP_var_indicios_manuales2
       Num_trimestres_en_deuda       antiguedad     indice_beneficio
```

```
        media_CLASIFICACION_CIF1      media_beneficio       prob_local     sector_v2
        trimestre      variabilidad_beneficio/ level=int;
;
input CLASIFICACION_CIF6_v2 CLASIFICACION_CIF5_v2 Tener_estimado_fraude_grupo
estm_fraude_grup_trim_v2 / level=nom
;
target  fraude;
hidden 1 / act=TAN id=H;
train tech=levmar;
score data=datos.validacion out=resumen.salpredi outfit=resumen.salfit;
run;


/*--------------------RANDOM FOREST----------------------*/


        %macro randomforest(vardep=,listconti=,listcategor=,
maxtrees=,variables=,porcenbag=,numvariables=,maxbranch=,tamhoja=,maxdepth=,pvalor=);

/* division train test */
 /*proc printto log='null'; RUN;*/
ods listing close;
proc hpforest data=datos.forest
maxtrees=&maxtrees
vars to try=&variables
trainfraction=0.6
leafsize=&tamhoja
maxdepth=&maxdepth
alpha=&pvalor
exhaustive=5000
missing=useinsearch ;
target &vardep/level=nominal;
input &listconti/level=interval;
%if (&listcategor ne) %then %do;
        input &listcategor/level=nominal;
%end;
ods output fitstatistics=resumen.forest_fitstatistics
variableimportance=resumen.forest_variableimportance;
score out=resumen.salpredi_forest;
/*save file="C:\Users\DepEIOIII\Desktop\model15.bin";*/

run;

data resumen.salpredi_forest2; set resumen.salpredi_forest; run;

/*proc hp4score data=&ArchivoValidacion;
id &listconti
   &listcategor
   &vardep;
score file="C:\Users\DepEIOIII\Desktop\model15.bin" out=salo;
run;*/
ods listing ;

data resumen.salpredi_forest; set resumen.salpredi_forest; if fraude=. then output; run;
data resumen.salpredi_forest; set resumen.salpredi_forest; id=_N_;run;
data resumen.salpredi_forest (keep=P_fraude1 id); set resumen.salpredi_forest;run;
data resumen.salpredi_forest (keep=fraude P_fraude1 id); merge resumen.salpredi_forest
datos.validacion;by id; run;

%do cortemin=1 %to 41 %by 1;
/*%do cortemax=7 %to 7 %by 1;*/
        proc printto log='null'; RUN;
        data tabla1 (keep=fraude P fraude1  PuntoDeCorteMin ); set
resumen.salpredi_forest; PuntoDeCorteMin=0.5+(&corteMin-1)*0.01; run;
        data tabla2; set tabla1; if (P_fraude1>=PuntoDeCorteMin) then fraude_estimado=1;
else fraude_estimado=-1; run;

        proc sql noprint; create table VP as select count(*) as VP from tabla2 where
(fraude=-1 and fraude_estimado=-1);quit;
```

```
        proc sql noprint; create table FP as select count(*) as FP from tabla2 where
(fraude=-1 and fraude_estimado=1);quit;
        proc sql noprint; create table FN as select count(*) as FN from tabla2 where
(fraude=1 and fraude_estimado=-1);quit;
        proc sql noprint; create table VN as select count(*) as VN from tabla2 where
(fraude=1 and fraude_estimado=1);quit;
        proc sql noprint; create table Rules as select sum(NRules) as Numrules from
resumen.forest_variableimportance ;run;


               data criterio; merge VP FP FN VN Rules; run;
               data criterio; set criterio;
               Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));
               PuntoDeCorteMin=0.51+(&corteMin-1)*0.01;
        maxtrees=&maxtrees ;
        vars_to_try=&variables;
        trainfraction=&porcenbag;
        leafsize=&tamhoja;
        maxdepth=&maxdepth;
        alpha=&pvalor;
        exhaustive=5000;
               numvariables=&numvariables;
               run;
               proc printto log=LOG; RUN;
               data resumen.resumen_forest; set resumen.resumen_forest criterio;run;
/*%END;*/
%end;
%mend;


options mprint=0;options notes;


%macro Forest;
%let lista2='0.1 0.05 0.01';
%let nume2=3;
%let lista3='1000 700 500 100';
%let nume3=4;
%let lista4='10 50 100';
%let nume4=3;
%do  j=1 %to &nume2 %by 1;
data  null ;p valor=scanq(&lista2,&j);call symput('p_valor',left(p_valor));run;
%do  k=1 %to &nume3 %by 1;
data  null ;tamhoja=scanq(&lista3,&k);call symput('tamhoja',left(tamhoja));run;
%do  l=1 %to &nume4 %by 1;
data  null ;maxtrees=scanq(&lista4,&l);call symput('maxtrees',left(maxtrees));run;
%do variables=5 %to 25 %by 5; /*5*/
%do tamhoja=100 %to 2100 %by 500; /*5*/
%do maxdepth=25 %to 25 %by 1; /*El maximo es 50*/
%do maxbranch=2 %to 2 %by 2; /*5*/

%randomforest(
vardep=fraude,
listconti=
F1      F2      F4      F6      F7      F8      Factor1 MP_CLASIFICACION_CIF3
        MP_CLASIFICACION_CIF7 MP_CLASIFICACION_CIF8
    MP_CLASIFICACION_CIF9     MP_DefinitivaVSProvisional    MP_Var_indicios_aut2
        MP_Var_indicios_aut3  MP_Var_indicios_aut4
    MP_Var_indicios_manuales1 MP_estado_reclamo_refact      MP_medio_reclamo_refact
        MP_salto_beneficio    MP_var_indicios_manuales2
    Num_trimestres_en_deuda   antiguedad     indice_beneficio
        media_CLASIFICACION_CIF1      media_beneficio       prob_local     sector_v2
    trimestre  variabilidad_beneficio
,
listcategor= CLASIFICACION_CIF6_V2 TENER_ESTIMADO_FRAUDE_GRUPO CLASIFICACION_CIF5_V2
ESTM FRAUDE GRUP TRIM V2,
maxtrees=&maxtrees.,variables=&variables.,porcenbag=0.6,numvariables=&variables.,maxbran
ch=&maxbranch.,tamhoja=&tamhoja.,maxdepth=&maxdepth.,pvalor=&p_valor.);
%end;
%end;
```

```
%end;
%end;
%end;
%end;
%end;
%mend;

%forest;

libname resumen "C:\Users\Dani\Dropbox\Master\TFM\Resumenes\Ramdon Forest";run;

data resumen.resumen_forest; set resumen.resumen_forest;if _N_=1 then delete ;run;
data resumen.resumen_forest (drop=exhaustive vars_to_try trainfraction); set
resumen.resumen_forest;run;



/*GRADIENT BOOSTING*/
libname datos "C:\Users\Dani\Desktop\AaTFM\Datos TFM\Muestra";run;
libname resumen "C:\Users\Dani\Desktop\AaTFM\Datos TFM\Totales\resumenes\Boosting";run;

data resumen.resumen_boosting;run;
proc printo
%macro boosting;
%let lista1='2100 1600 1100 600 100';
%let nume1=5;
%let lista2='100 50 10';
%let nume2=3;
%do  k=1 %to &nume1 %by 1;
data _null_;tamhoja=scanq(&lista1,&k);call symput('tamhoja',left(tamhoja));run;
%do  l=1 %to &nume2 %by 1;
data  null ;maxtrees=scanq(&lista2,&l);call symput('maxtrees',left(maxtrees));run;
%do maxdepth=25 %to 25 %by 1; /*El maximo es 50*/
%do maxbranch=2 %to 2 %by 2; /*5*/
%let lista3='0.01 0.05 0.09 1.3 1.7';
%let nume3=5;
%do  m=1 %to &nume3 %by 1;
data _null_;shrink=scanq(&lista3,&m);call symput('shrink',left(shrink));run;
%let lista4='10 50 100';
%let nume4=3;
%do  n=1 %to &nume4 %by 1;
data _null_;iterations=scanq(&lista4,&n);call symput('iterations',left(iterations));run;

proc treeboost data=datos.aprendizaje shrinkage=&shrink maxbranch=&maxbranch
maxdepth=&maxdepth iterations=&iterations leafsize=&tamhoja;
       input CLASIFICACION_CIF6_V2 TENER_ESTIMADO_FRAUDE_GRUPO CLASIFICACION_CIF5_V2
ESTM_FRAUDE_GRUP_TRIM_V2/level=nominal;
       input F1       F2      F4      F6      F7      F8      Factor1 MP_CLASIFICACION_CIF3
       MP_CLASIFICACION_CIF7 MP_CLASIFICACION_CIF8
    MP_CLASIFICACION_CIF9     MP_DefinitivaVSProvisional    MP_Var_indicios_aut2
       MP_Var_indicios_aut3   MP_Var_indicios_aut4
    MP_Var_indicios_manuales1 MP_estado_reclamo_refact      MP_medio_reclamo_refact
       MP_salto_beneficio     MP_var_indicios_manuales2
    Num_trimestres_en_deuda   antiguedad     indice_beneficio
       media_CLASIFICACION_CIF1      media_beneficio       prob_local      sector_v2
    trimestre  variabilidad_beneficio/level=interval;
       target fraude /level=nominal;
       SAVE FIT=resumen.FIT IMPORTANCE=resumen.IMP MODEL=resumen.MDL
RULES=resumen.RULES;
       score data=datos.validacion out=resumen.salpredi_boosting;
run;
ods listing ;


%do cortemin=1 %to 41 %by 1;
        data tabla1 (keep=fraude P_fraude1  PuntoDeCorteMin ); set
resumen.salpredi_boosting; PuntoDeCorteMin=0.5+(&corteMin-1)*0.01; run;
```

```
        data tabla2; set tabla1; if (P_fraude1>=PuntoDeCorteMin) then fraude_estimado=1;
else fraude_estimado=-1; run;

        proc sql noprint; create table VP as select count(*) as VP from tabla2 where
(fraude=-1 and fraude_estimado=-1);quit;
        proc sql noprint; create table FP as select count(*) as FP from tabla2 where
(fraude=-1 and fraude_estimado=1);quit;
        proc sql noprint; create table FN as select count(*) as FN from tabla2 where
(fraude=1 and fraude_estimado=-1);quit;
        proc sql noprint; create table VN as select count(*) as VN from tabla2 where
(fraude=1 and fraude estimado=1);quit;
        proc sql noprint; create table Rules as select sum(NRules) as Numrules from
resumen.IMP;run;


                data criterio; merge VP FP FN VN Rules; run;
                data criterio; set criterio;
                Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));
                PuntoDeCorteMin=0.51+(&corteMin-1)*0.01;
        maxtrees=&maxtrees ;
        leafsize=&tamhoja;
        maxdepth=&maxdepth;
        exhaustive=5000;
                shrinkage=&shrink;
                iterations=&iterations;
                maxbranch=&maxbranch;
                run;
                data resumen.resumen_boosting; set resumen.resumen_boosting criterio;run;
%end;
%end;
%end;
%end;
%end;
%end;
%end;
%mend;

%boosting;


/*----Ensamble de Modelos----*/
/*----librerias---*/
libname datos "C:\Users\Dani\Desktop\AaTFM\Datos TFM\Totales";run;
libname resumen "C:\Users\Dani\Desktop\ensamble";run;
libname pred "C:\Users\Dani\Dropbox\Master\TFM\Resumenes\Predicciones";run;

/*---preparacion de los datos----*/
data pred.fraude (keep=fraude);set datos.validacion;run;
data pred.salpredi_logistica (keep=P_1);set pred.salpredi_logistica; run; data
pred.salpredi logistica (rename=(P 1=P1 logistica)); set pred.salpredi logistica;run;
data pred.salpredi_redes (keep=P_fraude1);set pred.salpredi_redes;run; data
pred.salpredi_redes (rename=(P_fraude1=P1_redes)); set pred.salpredi_redes;run;
data pred.salpredi_forest (keep=P_fraude1);set pred.salpredi_forest;run; data
pred.salpredi_forest (rename=(P_fraude1=P1_forest)); set pred.salpredi_forest;run;
data pred.salpredi_boosting (keep=P_fraude1);set pred.salpredi_boosting;run; data
pred.salpredi_boosting (rename=(P_fraude1=P1_boosting)); set pred.salpredi_boosting;run;
data pred.predicciones; merge pred.salpredi_logistica pred.salpredi_redes
pred.salpredi_forest pred.salpredi_boosting pred.fraude;run;


data resumen.Resumen_ensamble;run;

/*------------------------orden 2--------------------------------*/
/*---logistica/redes----*/
%macro ensamble1;
%let lista1='min max median mean';
%let nume1=4;
%do  i=1 %to &nume1 %by 1;
data _null_;stad=scanq(&lista1,&i);call symput('stad',left(stad));run;
```

```
data tabla1 (keep=P1_logistica P1_redes fraude P1); set
pred.predicciones;P1=&stad(P1_logistica, P1_redes); run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude_estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude_estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude_estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=1;
Redes=1;
Forest=0;
Boosting=0;
Agregacion="&stad";
p_corte=&corte/10;
run;
data resumen.Resumen_ensamble; set resumen.Resumen_ensamble criterio;run;
%end;
%end;
%mend;


/*---logistica/forest----*/
%macro ensamble2;
%let lista1='min max median mean';
%let nume1=4;
%do  i=1 %to &nume1 %by 1;
data _null_;stad=scanq(&lista1,&i);call symput('stad',left(stad));run;
data tabla1 (keep=P1_logistica P1_forest fraude P1); set
pred.predicciones;P1=&stad(P1_logistica, P1_forest); run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude_estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude_estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude_estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=1;
redes=0;
Forest=1;
Boosting=0;
Agregacion="&stad";
p_corte=&corte/10;
run;
data resumen.Resumen_ensamble; set resumen.Resumen_ensamble criterio;run;
%end;
%end;
%mend;

/*---logistica/boosting----*/
%macro ensamble3;
%let lista1='min max median mean';
%let nume1=4;
%do  i=1 %to &nume1 %by 1;
```

```
data _null_;stad=scanq(&lista1,&i);call symput('stad',left(stad));run;
data tabla1 (keep=P1_logistica P1_boosting fraude P1); set
pred.predicciones;P1=&stad(P1_logistica, P1_boosting); run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude_estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude_estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=1;
redes=0;
Forest=0;
Boosting=1;
Agregacion="&stad";
p_corte=&corte/10;
run;
data resumen.Resumen_ensamble; set resumen.Resumen_ensamble criterio;run;
%end;
%end;
%mend;


/*---forest/redes----*/
%macro ensamble4;
%let lista1='min max median mean';
%let nume1=4;
%do  i=1 %to &nume1 %by 1;
data _null_;stad=scanq(&lista1,&i);call symput('stad',left(stad));run;
data tabla1 (keep=P1_forest P1_redes fraude P1); set
pred.predicciones;P1=&stad(P1_forest, P1_redes); run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude_estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude_estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=0;
redes=1;
Forest=1;
Boosting=0;
Agregacion="&stad";
p_corte=&corte/10;
run;
data resumen.Resumen_ensamble; set resumen.Resumen_ensamble criterio;run;
%end;
%end;
%mend;

/*---boosting/redes----*/
%macro ensamble5;
%let lista1='min max median mean';
%let nume1=4;
```

```
%do  i=1 %to &nume1 %by 1;
data _null_;stad=scanq(&lista1,&i);call symput('stad',left(stad));run;
data tabla1 (keep=P1_boosting P1_redes fraude P1); set
pred.predicciones;P1=&stad(P1_boosting, P1_redes); run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude_estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude_estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=0;
redes=1;
Forest=0;
Boosting=1;
Agregacion="&stad";
p_corte=&corte/10;
run;
data resumen.Resumen_ensamble; set resumen.Resumen_ensamble criterio;run;
%end;
%end;
%mend;

/*---forest/boosting----*/
%macro ensamble6;
%let lista1='min max median mean';
%let nume1=4;
%do  i=1 %to &nume1 %by 1;
data _null_;stad=scanq(&lista1,&i);call symput('stad',left(stad));run;
data tabla1 (keep=P1_forest P1_boosting fraude P1); set
pred.predicciones;P1=&stad(P1_forest, P1_boosting); run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude_estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude_estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=0;
redes=0;
Forest=1;
Boosting=1;
Agregacion="&stad";
p_corte=&corte/10;
run;
data resumen.Resumen_ensamble; set resumen.Resumen_ensamble criterio;run;
%end;
%end;
%mend;

/*------------------------orden 3--------------------------------*/
/*---logistica/redes/forest----*/

%macro ensamble7;
```

```
%let lista1='min max median mean';
%let nume1=4;
%do  i=1 %to &nume1 %by 1;
data _null_;stad=scanq(&lista1,&i);call symput('stad',left(stad));run;
data tabla1 (keep=P1_logistica P1_redes P1_forest fraude P1); set
pred.predicciones;P1=&stad(P1_logistica, P1_redes, P1_forest); run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude_estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude_estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=1;
redes=1;
Forest=1;
Boosting=0;
Agregacion="&stad";
p_corte=&corte/10;
run;
data resumen.Resumen_ensamble; set resumen.Resumen_ensamble criterio;run;
%end;
%end;
%mend;

/*---logistica/redes/boosting----*/

%macro ensamble8;
%let lista1='min max median mean';
%let nume1=4;
%do  i=1 %to &nume1 %by 1;
data _null_;stad=scanq(&lista1,&i);call symput('stad',left(stad));run;
data tabla1 (keep=P1_logistica P1_redes P1_boosting fraude P1); set
pred.predicciones;P1=&stad(P1_logistica, P1_redes, P1_boosting); run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude_estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude_estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude_estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=1;
redes=1;
Forest=0;
Boosting=1;
Agregacion="&stad";
p_corte=&corte/10;
run;
data resumen.Resumen_ensamble; set resumen.Resumen_ensamble criterio;run;
%end;
%end;
%mend;

/*---forest/redes/boosting----*/
```

```
%macro ensamble9;
%let lista1='min max median mean';
%let nume1=4;
%do  i=1 %to &nume1 %by 1;
data _null_;stad=scanq(&lista1,&i);call symput('stad',left(stad));run;
data tabla1 (keep=P1_forest P1_redes P1_boosting fraude P1); set
pred.predicciones;P1=&stad(P1_forest, P1_redes, P1_boosting); run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude_estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude_estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=0;
redes=1;
Forest=1;
Boosting=1;
Agregacion="&stad";
p_corte=&corte/10;
run;
data resumen.Resumen_ensamble; set resumen.Resumen_ensamble criterio;run;
%end;
%end;
%mend;

/*---forest/logistica/boosting----*/

%macro ensamble10;
%let lista1='min max median mean';
%let nume1=4;
%do  i=1 %to &nume1 %by 1;
data _null_;stad=scanq(&lista1,&i);call symput('stad',left(stad));run;
data tabla1 (keep=P1_forest P1_logistica P1_boosting fraude P1); set
pred.predicciones;P1=&stad(P1_forest, P1_logistica, P1_boosting); run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude_estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude_estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude_estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=1;
redes=0;
Forest=1;
Boosting=1;
Agregacion="&stad";
p_corte=&corte/10;
run;
data resumen.Resumen_ensamble; set resumen.Resumen_ensamble criterio;run;
%end;
%end;
%mend;
```

```
/*-----------------------orden 4--------------------------------*/
/*---logistica/redes/forest/boosting----*/

%macro ensamble11;
%let lista1='min max median mean';
%let nume1=4;
%do  i=1 %to &nume1 %by 1;
data _null_;stad=scanq(&lista1,&i);call symput('stad',left(stad));run;
data tabla1 (keep=P1_logistica P1_redes P1_forest P1_boosting fraude P1); set
pred.predicciones;P1=&stad(P1_logistica, P1_redes, P1_forest, P1_boosting); run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude_estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude_estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude_estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=1;
redes=1;
Forest=1;
Boosting=1;
Agregacion="&stad";
p_corte=&corte/10;
run;
data resumen.Resumen_ensamble; set resumen.Resumen_ensamble criterio;run;
%end;
%end;
%mend;

%ensamble1;
%ensamble2;
%ensamble3;
%ensamble4;
%ensamble5;
%ensamble6;
%ensamble7;
%ensamble8;
%ensamble9;
%ensamble10;
%ensamble11;


proc sort data=resumen.Resumen_ensamble; by descending matthews; run;



/*-----------------Ensambles con regresion-----*/
libname datos "C:\Users\Dani\Desktop\AaTFM\Datos TFM\Totales";run;
libname resumen "C:\Users\Dani\Desktop\ensamble";run;
libname pred "C:\Users\Dani\Desktop\ensamble\Predicciones_aprendizaje";run;
libname pred2 "C:\Users\Dani\Dropbox\Master\TFM\Predicciones\Validacion";run;
libname pred3 "C:\Users\Dani\Dropbox\Master\TFM\Predicciones\Test";run;
/*---preparacion de los datos----*/

data pred.fraude (keep=fraude);set datos.aprendizaje;run;
data pred.salpredi_logistica (keep=P_1);set pred.salpredi_logistica; run; data
pred.salpredi_logistica (rename=(P_1=P1_logistica)); set pred.salpredi_logistica;run;
data pred.salpredi redes (keep=P fraude1);set pred.salpredi redes;run; data
pred.salpredi_redes (rename=(P_fraude1=P1_redes)); set pred.salpredi_redes;run;
data pred.salpredi_forest (keep=P_fraude1);set pred.salpredi_forest;run; data
pred.salpredi_forest (rename=(P_fraude1=P1_forest)); set pred.salpredi_forest;run;
```

```
data pred.salpredi_boosting (keep=P_fraude1);set pred.salpredi_boosting;run; data
pred.salpredi_boosting (rename=(P_fraude1=P1_boosting)); set pred.salpredi_boosting;run;
data pred.predicciones; merge pred.salpredi_logistica pred.salpredi_redes
pred.salpredi_forest pred.salpredi_boosting pred.fraude;run;

/*----Logistica---*/
proc logistic data=pred.predicciones ;
    model fraude (event='1')= P1_logistica P1_redes P1_forest P1_boosting/noint

       selection=stepwise  SLENTRY=0.05 parmlabel rsquare lackfit clparm=wald clodds=pl
     ctable pprob = (.05 to .95 by .05);
    score data=pred.predicciones out=pred.salpredi_ensamble1;
       score data=pred2.predicciones out=pred2.salpredi_ensamble2;
       /*score data=pred3.test out=pred3.salpredi_logistica_t;*/
    run;

data resumen.ensamble1;run;
%macro ensamble_logistica1;
data pred2.salpredi_ensamble1 (keep=P_1 fraude);set pred2.salpredi_ensamble1; run; data
pred2.salpredi_ensamble1 (rename=(P_1=P1)); set pred2.salpredi_ensamble1;run;
data tabla1; set pred2.salpredi_ensamble1;run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude_estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude_estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude_estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
p_corte=&corte/10;
run;
data resumen.ensamble1; set resumen.ensamble1 criterio;run;
%end;
%mend;

%ensamble_logistica1;


data resumen.ensamble2;run;
%macro ensamble2;

data tabla1 (keep=P1 logistica P1 redes P1 forest P1 boosting fraude P1); set
pred2.predicciones;P1=P1 logistica*(7.8101/31.0901)+ P1 redes*(4.4647/31.0901)+
P1_forest*(3.8744/31.0901)+ P1_boosting*(14.9909/31.0901); run;
%do corte=4 %to 10 %by 1;
data tabla2; set tabla1; if (P1>&corte/10) then fraude_estimado=1; else
fraude_estimado=-1; run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude_estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude_estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=1;
redes=1;
Forest=1;
Boosting=1;
```

```
Agregacion="&stad";
p_corte=&corte/10;
run;
data resumen.ensamble2; set resumen.ensamble2 criterio;run;
%end;
%mend;

%ensamble2;

proc sort data=resumen.ensamble1; by descending matthews; run;
proc sort data=resumen.ensamble2; by descending matthews; run;

proc export data=resumen.ensamble1
outfile="C:\Users\Dani\Dropbox\Master\TFM\salidas\ensamble1.xls"  dbms=EXCEL2007;
run;

proc export data=resumen.ensamble2
outfile="C:\Users\Dani\Dropbox\Master\TFM\salidas\ensamble2.xls"  dbms=EXCEL2007;
run;



/*-----------------datos test-------*/

libname pred3 "C:\Users\Dani\Dropbox\Master\TFM\Predicciones\Test";run;


/*---preparacion de los datos----*/
data pred3.fraude (keep=fraude);set datos.test;run;
data pred3.salpredi_redes (keep=P_fraude1);set pred3.salpredi_redes;run; data
pred3.salpredi redes (rename=(P fraude1=P1 redes)); set pred3.salpredi redes;run;
data pred3.salpredi_boosting (keep=P_fraude1);set pred3.salpredi_boosting;run; data
pred3.salpredi boosting (rename=(P_fraude1=P1_boosting)); set
pred3.salpredi_boosting;run;
data pred3.predicciones; merge  pred3.salpredi_redes  pred3.salpredi_boosting
pred3.fraude;run;

/*---boosting/redes----*/
libname ressumen "C:\Users\Dani\Dropbox\Master\TFM\Resumenes\Estimacion Insesgada"; run;
data resumen.fin; run;
data tabla1 (keep=P1_boosting P1_redes fraude P1); set
pred3.predicciones;P1=mean(P1_boosting, P1_redes); run;

data tabla2; set tabla1; if (P1>=0.5) then fraude_estimado=1; else fraude_estimado=-1;
run;
proc sql noprint; create table VP as select count(*) as VP from tabla2 where (fraude=-1
and fraude_estimado=-1);quit;
proc sql noprint; create table FP as select count(*) as FP from tabla2 where (fraude=-1
and fraude estimado=1);quit;
proc sql noprint; create table FN as select count(*) as FN from tabla2 where (fraude=1
and fraude_estimado=-1);quit;
proc sql noprint; create table VN as select count(*) as VN from tabla2 where (fraude=1
and fraude_estimado=1);quit;
data criterio; merge VP FP FN VN; run;
data criterio; set criterio;
Matthews=((VP*VN)-(FP*FN))/(sqrt((VP+FP)*(VP+FN)*(VN+FP)*(VN+FN)));run;
data criterio;set criterio;
Logistica=0;
redes=1;
Forest=0;
Boosting=1;
Agregacion="Media/Mediana";
p_corte=0.5;
run;
data resumen.fin; set resumen.fin criterio;run;
```

12. Bibliografía.

12.1. Bibliografia referenciada en el texto del presente estudio.

1. Daniel Peña, Análisis de datos multivariante.
2. Eduardo Francisco Caicedo Bravo, Aproximación practica a las Redes Neuronales Artificiales.
3. Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A:1010933404324.7
4. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". Biochimica et Biophysica Acta (BBA) - Protein Structure 405 (2): 442–451. doi:10.1016/0005-2795(75)90109-9.
5. Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation"
6. https://en.wikipedia.org/wiki/Early_stopping.
7. http://www.dm.uba.ar/materias/modelos_lineales_generalizados_Mae/2005/1/glm5.pdf.
8. http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/CUALITATIVAS/LOGISTICA/regresion-logistica.pdf.
9. https://es.wikipedia.org/wiki/Funci%C3%B3n_de_activaci%C3%B3n

12.2. Resto de bibliografia utilizada.

Aproximación practica a las Redes Neuronales Artificiales / Eduardo Francisco Caicedo Bravo.

Análisis multivariante I / José Luis Valencia Delfa, Maria Lina Vicente Hernanz.

Técnicas de análisis multivariante / César Pérez López.

Técnicas de Análisis Multivariante de Datos: aplicaciones con SPSS / César Pérez López

Minería de datos. Redes Neuronales y Árboles de decisión: Ejemplos con SAS Enterprise Miner / César Pérez López.

Aplicación de las redes neuronales artificiales a la regresión / Quintin Martín Martín, Yanira del Rosario de Paz Santana.