UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES ACATLÁN

EL USO DE LA MINERÍA DE DATOS COMO UNA HERRAMIENTA PARA LA TOMA DE DECISIONES EN LA COBRANZA Y LAS DEVOLUCIONES DE IMPUESTOS DENTRO DEL SERVICIO DE ADMINISTRACIÓN TRIBUTARIA (SAT)

TESINA

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN MATEMÁTICAS APLICADAS Y COMPUTACIÓN

PRESENTA

LORENZO LÓPEZ AGUILAR

Asesor: SARA CAMACHO CANCINO

Abril 2015

Santa Cruz Acatlán, Estado de México

Dedicatoria.

Dedico este trabajo a:

Mis padres. Anita y Lorenzo, muchas gracias por todo, los amo.

A mi esposa. Lolis, gracias por tu amor, apoyo, comprensión y tolerancia, te amo.

A mi hija. Anita Lu, hermosa, debes terminar todo lo que empiezas, te amo.

A mi familia. Gracias por estar siempre ahí, me siento querido.

A mis amigos. Gracias por su amistad, la valoro mucho.

A la UNAM. Gracias a nuestra máxima casa de estudios muchos hemos tenido la oportunidad de aprender y ser profesionistas.

A mis profesores. Gracias por querer compartir sus conocimientos.

A mi asesora. Sara, muchas gracias por tu gran vocación, conocimiento, orientación, tiempo y apoyo.

A todos los que siempre creyeron en mí. Muchas gracias.

Índice.

Introducción…… 1

Capítulo I. La minería de datos…… 3
I.1. ¿Qué es la minería de datos?..... 3
I.2. Aprendizaje supervisado…… 4
I.3. Aprendizaje no supervisado…….. 32
I.4. Recapitulación……. 40

Capítulo II. Las devoluciones y la cobranza en el SAT…… 41
II.1. Las devoluciones…… 45
II.2. La cobranza……. 55
II.3. Recapitulación……. 71

Capítulo III. Entendimiento y preparación de datos……. 72
III.1. Entendimiento de los datos……. 75
III.2. Preparación de los datos……. 94
III.3. Recapitulación…… 110

Capítulo IV. Modelado y evaluación…… 111
IV.1. Devoluciones……. 112
IV.2. Cobranza…… 135
IV.3. Ciclo de vida de los modelos…….. 155
IV.4. Recapitulación……... 156

Conclusiones…….. 157

Bibliografía…….. 159

Introducción.

En la vida diaria siempre estamos tomando decisiones, en la mayoría de los casos con base en la experiencia, decisiones como: si se debe atravesar una calle, abrigarse o no de acuerdo a condiciones climáticas, elegir entre determinada ropa para una fiesta en un lugar específico.

La experiencia nos hace tomar decisiones para realizar las acciones que creemos serán las adecuadas en una situación específica.

Dentro de las organizaciones la toma de decisiones es una etapa muy importante en las actividades que se realizan, la información que se tiene para realizarlas es fundamental ya que podría determinar el éxito o el fracaso de un proyecto.

Una organización en su operación diaria está generando información con sus diferentes sistemas transaccionales, ésta puede ser almacenada en distintos repositorios, estructurados y no estructurados. En la mayoría de los casos se utiliza como consultas, reportes o cubos de información.

Este tipo de análisis es únicamente descriptivo, muestra un comportamiento histórico de las transacciones, se generan reportes de hechos pasados, sin embargo también puede utilizarse para encontrar tendencias de comportamiento que resulten en predicciones con un grado de precisión para responder preguntas a determinados problemas.

Los modelos predictivos son herramientas muy importantes que mediante información histórica pueden generar pronósticos a futuro del comportamiento de un problema que se intenta resolver.

Al generar modelos predictivos se está estimando la probabilidad de ocurrencia de un evento, esto da una gran ventaja, con información descriptiva y predictiva una organización puede tomar decisiones no sólo para analizar el presente, sino para planear a futuro.

El desarrollo del trabajo mostrará como elaborar modelos predictivos usando la minería de datos para generar pronósticos en dos procesos dentro de la administración tributaria: la cobranza y las devoluciones.

Se eligieron estos dos temas porque son formas de ingresos y egresos que se puede tener en una organización para su operación y mostrará cómo es que los modelos predictivos pueden ayudar en la toma de decisiones.

En el capítulo uno se dará un resumen de las técnicas minería de datos tanto para la predicción como para describir los datos. No se pretende un análisis exhaustivo ni completo sino dar una idea general de las técnicas más usadas.

En el segundo capítulo se abordará de manera general lo que son los procesos de cobranza y devoluciones de impuestos dentro del Servicio de Administración Tributaria.

El capítulo tres mostrará la forma de extraer, trasformar y cargar la información con el fin de prepararla para el modelado.

En el capítulo cuatro se generarán modelos predictivos y se evaluarán para los dos procesos en estudio. Para el tema de la cobranza se construirán modelos predictivos con los cuales se mostrará cómo pueden ayudar a determinar la estimación de probabilidad de pago de un adeudo así como generar estrategias de cobro.

Para el caso de las devoluciones, se mostrará cómo los modelos pueden ayudar en la resolución de trámites determinando el grado de riesgo de acuerdo a la historia de solicitudes ya resueltas.

Al final se darán las conclusiones de acuerdo a los resultados obtenidos.

.

Capítulo I. La minería de datos.

I.1. ¿Qué es la minería de datos?

La minería de datos es la extracción de conocimiento útil y novedoso en grandes volúmenes de información mediante modelos descriptivos o predictivos que permitan la toma de decisiones.

Toda organización tiene información almacenada que es proporcionada por los sistemas transaccionales con los que trabaja diario, los cuales permiten la operación normal, esta información se guarda en almacenes de datos, éstos pueden ser estructurados o no estructurados.

Los sistemas transaccionales son aquellos que permiten la operación de una organización, por ejemplo los sistemas contables, de recursos humanos, de registro y control, de finanzas etc., éstos en las organizaciones generan información histórica que con el tiempo se vuelve muy grande, ésta puede servir para sistemas de consulta.

La información que generan los sistemas transaccionales se almacena en estructuras de datos (también pueden ser no estructurados, como documentos de texto) que permiten su consulta posterior, mediante herramientas de consulta como SQL (Structured Query Language, Lenguaje Estructurado de Consulta), cubos o sistemas de consulta se pueden generar reportes para presentar los datos almacenados.

En organizaciones grandes la información es inmensa, muchos datos organizados en distintas estructuras de datos que tienen información valiosa desconocida. La estadística es la disciplina utilizada tradicionalmente para obtener conocimiento de los datos, mediante técnicas matemáticas nos ofrece tanto descripciones de los datos como análisis y predicciones de los mismos.

La minería de datos es una disciplina muy flexible en el manejo de grandes volúmenes de información tanto en número de casos (observaciones, renglones o registros) como en número de atributos (campos o columnas), tiene diferentes técnicas que se adaptan a las necesidades de conocimiento deseado y su fundamento está en el aprendizaje automático y la inteligencia artificial. Los datos que se usarán en la minería de datos deben ser tipificados (identificar el tipo de dato) para poder generar modelos, pueden ser datos de tipo:

Rangos: Variables numéricas (como por ejemplo Ingresos) éstas tienen un tratamiento en base a sus valores.

Conjunto: Variables numéricas o de cadena que representan grupos (por ejemplo los colores de algún objeto) o números que representan una categoría (por ejemplo 1 Guanajuato, 2, Jalisco, 3 Tamaulipas, etc.) éstas son utilizadas como conjunto de datos.

Dicotómicas: Variables con solo dos valores posibles numéricos o de cadena.

Las dos grandes ramas de la minería de datos son el aprendizaje supervisado y el no supervisado, el primero trata de predecir al futuro basándose en el pasado, hace estimaciones de datos conocidos y resueltos generando modelos que aplicará a datos nuevos para tener un pronóstico, el segundo describe los datos basándose en patrones, encontrando relaciones entre ellos generando modelos que permite conocer su comportamiento.

I.2. Aprendizaje supervisado.

El fundamento del aprendizaje supervisado es aprender del pasado para predecir el futuro generando modelos predictivos de minería de datos. Utilizando una muestra de información a analizar suficiente (normalmente se necesita una historia de al menos 5 años, dependiendo de la naturaleza de la misma podría ser menos) de casos conocidos, ésta será la historia de la cual el modelo se entrenará, por eso el nombre de “supervisado”, se sabe que ocurrió.

El aprendizaje supervisado trata de responder la pregunta de qué es lo que pasará en el futuro de acuerdo a lo ocurrido en el pasado, cómo es que se han comportado los datos conocidos para conocer los desconocidos, encontrar patrones, conductas y pronósticos acerca del futuro, esto es muy importante ya que permite la toma oportuna de decisiones a priori permitiendo ser pro activos y no reactivos, además hace uso de los datos almacenados que en muchos de los casos solo sirve para sistemas de consulta y reportes

Necesita de dos clases de datos, los datos predictores y los predichos, hablando en términos matemáticos sería las variables independientes y las dependientes, en minería de datos se habla de las variables de entrada y la de salida (muchas veces conocida también como variable objetivo).

La variable de salida es la que nos va a dar el pronóstico de lo que pasará para casos nuevos, los casos conocidos nos indicarán lo que sucedió en el pasado tomando como aprendizaje las

variables de entrada, esto nos generará un modelo al que le podremos alimentar datos nuevos y nos dará una estimación de acuerdo a la variable de salida u objetivo.

Por ejemplo si hablamos de un banco con una historia de clientes clasificados por morosos o cumplidos podemos querer saber ¿cómo es que un cliente nuevo se comportará en el futuro?

La situación podría ser la siguiente, se tiene información de clientes ya conocida almacenada en una tabla histórica (Tabla I.1):

Edad	Ingresos	Antigüedad Trabajo	Morosidad
40	100,000	12	No
25	5,000	1	Si
39	30,000	10	No
22	3,000	2	Si
45	16,000	15	No
28	8,000	3	Si
31	18,000	5	Si
34	15,000	7	No
30	45,000	4	Si
26	9,000	6	No

Tabla I.1. Ejemplo información de clientes.

Tenemos las variables de entrada Edad, Ingresos, Antigüedad Trabajo, la variable de salida u objetivo es Morosidad la cual cuenta con solo dos valores (Si, No), la variable objetivo depende de las variables de entrada, cada uno de los renglones se les llama casos, por lo que se tienen 10 casos, el modelo podría generar un algoritmo como el siguiente:

1. Si Edad >= 39 y Ingresos >=16,000 y Antigüedad Trabajo >= 10 entonces Morosidad = No
2. Si Antigüedad Trabajo >=6 y Antigüedad Trabajo <=7 entonces Morosidad = No
3. De lo contrario Morosidad = Si

Cuando se quisiera saber si dos clientes nuevos serán morosos, se tienen los siguientes datos ficticios (Tabla I.2):

Edad	Ingresos	Antigüedad Trabajo	Morosidad
41	18,000	10	¿?
32	40,000	1	¿?

Tabla I.2. Ejemplo información de nuevos clientes.

Tenemos los datos de entrada, éstos los pasamos por el modelo generado previamente con la historia teniendo el siguiente resultado (Tabla I.3.)

Edad	Ingresos	Antigüedad Trabajo	Pronóstico Morosidad
41	18,000	10	No
32	40,000	1	Si

Tabla I.3. Ejemplo Pronóstico Morosidad.

Como podemos ver el modelo aplicó el algoritmo para los datos de entrada y obtuvo un pronóstico, el banco podría usar este resultado para saber si otorga un crédito.

Las diferentes técnicas de aprendizaje supervisado incluyen: árboles de decisión, redes neuronales, redes bayesianas y técnicas estadísticas; las cuales son explicadas en el resto del apartado I.2.

I.2.1. Árboles de decisión.

El nombre viene por la forma en la cual se dividen los datos en nodos (hojas) de un origen hacia otros nodos por las características de los datos (ramas), se van dividiendo conforme al algoritmo seleccionado, generando un modelo que en su forma gráfica se parece al de un árbol (Figura I.1.)

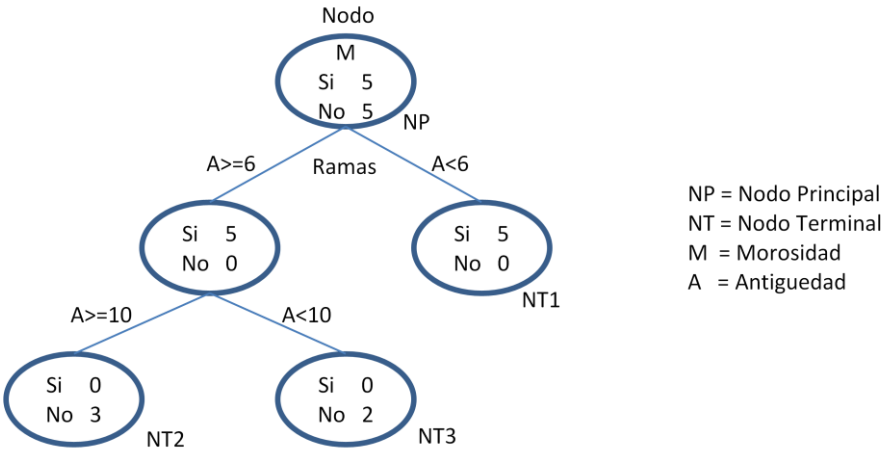


Figura I.1. Ejemplo árboles de decisión.

Inicia en un nodo principal el cual contiene la cantidad total de los casos con los cuales el modelo ha aprendido, a partir del principal se empieza dividir en diferentes nodos dependiendo de la variable por la cual el algoritmo empieza a generar el árbol, la primera variable que divide los datos es la más importante, esta importancia es calculada dependiendo del algoritmo usado (Chi cuadrada, entropía, etc), empieza a generar niveles los cuales tendrán profundidad (más niveles) y al final se tendrán los nodos terminales que es donde el algoritmo encuentra el pronóstico para esa rama.

La decisión final se puede encontrar siguiendo las ramas con sus condiciones que se generan desde el inicio del árbol (raíz) hasta alguno de sus nodos terminales (hojas finales). El algoritmo es uno de los más fáciles de interpretar dado su resultado en nodos con frecuencia de datos que pueden representarse con simples condiciones “si → entonces”, la salida gráfica permite la interpretación de manera muy simple e inteligible.

Como se puede observar el árbol clasifica los datos dependiendo de la condición generada, estas condiciones son excluyentes es decir pertenecen a una pero no a otra condición, no se puede pertenecer a ambas, a esto se le llama exhaustividad, lo que conducirá a una única ruta para cada ejemplo.

Lo más importante de los algoritmos de árboles de decisión es cómo elegir la partición para hacer crecer las ramas, es fundamental tener un criterio definido porque éste determinará la precisión de la clasificación.

Los principales criterios de selección de particiones se basan en la premisa de generar los nodos hijos más “puros”, es decir, cuando el número de casos generen una hoja con más valores que discriminen mejor la variable objetivo (Figura I.2), por ejemplo, se tienen la siguientes particiones:

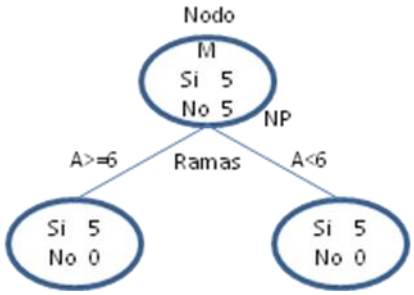


Figura I.2 Ejemplo nodos hijo puros.

Las particiones generadas son puras, ya que tienen 5 casos que representan Si y 5 que representan No para los dos nodos generados, esta es una partición óptima porque separa con exactitud los nodos resultantes y el crecimiento del árbol terminaría en ese nivel.

Por otro lado vemos el siguiente ejemplo:

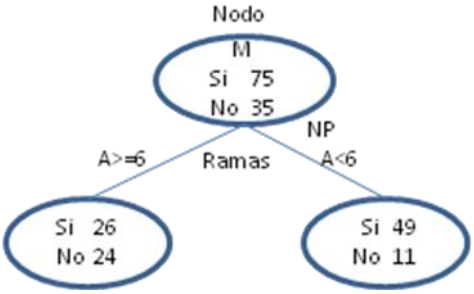


Figura I.2 Ejemplo nodos hijo no puros.

Se puede observar que el resultado de la partición no son nodos puros (Figura I.2), sin embargo el nodo A<6 tiene una discriminación de valores bastante clara, 49 casos para Si que representa el [49/(49+11)]*100=81.7% del total por lo tanto el 18.3% para No, podría ser un nodo terminal con bastante precisión para un caso nuevo.

Sin embargo se puede observar que el nodo A>=6 se tiene una cantidad de casos muy similar para los valores de la variable objetivo, por lo que un algoritmo debería de determinar si los nodos deben seguir creciendo para estos casos y tratarlos de hacer más claros discriminando la variable objetivo Si o No.

Los árboles de decisión también pueden utilizarse para la regresión, agrupamiento o estimación de probabilidades, en el ejemplo anterior se tiene una estimación de probabilidad de .817 para Si y un .183 para No, por lo que cualquier caso nuevo que siga esta rama y llegue al nodo terminal se le puede asignar la estimación mencionada.

Para estos ejemplos se está mostrando el caso más sencillo de un árbol binario, sin embargo también podría ser un árbol con más de dos nodos hijos que complicaría un poco más los resultados pero pudiera tener una mejor precisión de clasificación.

Los criterios de selección de particiones más conocidos son: Error esperado, el criterio GINI (CART), los criterios Gain (entropía), Gain Ratio y el DKM.

Es importante considerar en la construcción de un árbol de decisión que éste debe de ser lo suficientemente general para poderse adaptar a nuevos casos, si se tiene un árbol con una gran profundidad que genera una precisión muy alta con los datos de entrenamiento, podría ser no muy eficiente para nuevos datos, para eso existen técnicas de poda:

- Prepoda: Se realiza durante la construcción del árbol mediante un criterio de parada como el número de casos de la hoja, error esperado, etc.

- Pospoda: Se realiza después de la construcción de un árbol, eliminando nodos de las ramas hacia la raíz del árbol.

Los árboles de decisión también pueden generar conjuntos de reglas, éstas en general son más exhaustivas ya que presentan todos los caminos encontrados por el algoritmo, para el ejemplo ya visto:

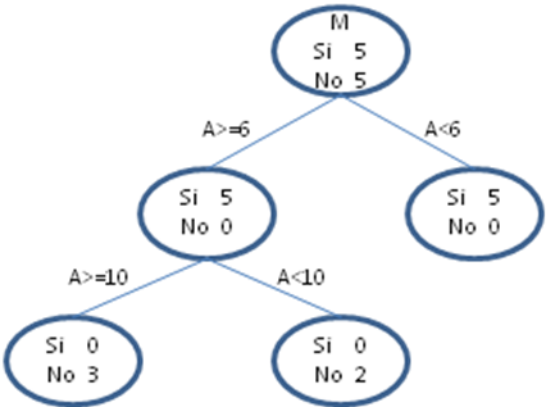


Figura I.3. Ejemplo árbol de decisión con sus nodos.

Un árbol de decisión generaría un algoritmo:

1. Si A<6 entonces Si
2. De lo contrario entonces No

En cambio un conjunto de reglas:

1. Si A<6 entonces Si
2. Si A>=6 y Si A>=10 entonces No
3. Si A>=6 y Si A<10 entonces No

El conjunto de reglas genera más condiciones y el árbol de decisión genera menos, esto para muchos más números de casos puede resultar en más simplicidad en el algoritmo de árbol que en el de reglas.

Los modelos de árbol más conocidos [1] son:

C&RT. (Clasification and Regression Trees) Genera particiones binarias. El algoritmo utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso. Los campos objetivo y predictor pueden ser de números o valores categóricos. Los mismos campos predictores pueden ser utilizados más de una vez en diferentes niveles del árbol. La poda se basa en la estimación de complejidad del error.

QUEST (Quick, Unbiased, Efficient Statistical Tree.) Genera particiones binarias para generar árboles de decisión; está diseñado para reducir el tiempo de procesamiento necesario para realizar los análisis de árboles de clasificación y regresión. Los campos predictores pueden ser categóricos o numéricos pero el campo objetivo debe ser categórico.

CHAID. (Chi-squared Automatic Interaction Detector.), Genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas, puede generar árboles no binarios. Los campos objetivo y predictor pueden ser de numéricos o categóricos.

C5.0. El algoritmo utiliza criterios de partición en base ganancia de la información (entropía), este árbol presume una mejora al original C4.5, utiliza la poda basada en reglas u otros algoritmos más sofisticados.

[1] Modelos de árbol de SPSS Clementine 12

Consideraciones sobre estos modelos.

Los árboles de decisiòn son algoritmos muy poderosos, que son inteligibles, esto los hace accesibles para ser explicados, en general cuando se les pone a competir contra otro tipo de algoritmos dan muy buenos resultdos en su precisión cuando se compara los valores correctamente clasificados con forme a la historia, este tipo de comparación se revisará en el capítulo IV donde se evaluarán diferentes tipos.

I.2.2. Redes neuronales.

Las redes neuronales son algoritmos que por su forma de trabajar tienen un cierto parecido en como las neuronas se comunican en el cerebro, de ahí su nombre, generando nodos de inicio que empiezan a interactuar para tener caminos de diferentes rutas que se entrelazan entre sí, empieza por una primer capa, esta es la capa principal de ahí dependiendo de la técnica a utilizar genera una o más capas ocultas para tener una capa terminal la cual tendrá los resultados finales.

Las redes neuronales tienen la ventaja de generar múltiples caminos por los cuales el flujo puede llegar al resultado, otra ventaja es que una red neuronal puede ser muy precisa debido a la cantidad de conexiones que genera, sin embargo puede ser tan particular que solo sirva para los datos con los cuales fue entrenada.

El esquema de rutas generadas por una red neuronal puede ser tan complicada que no es posible ver gráficamente los resultados como se puede hacer en otras técnicas de aprendizaje supervisado, esto hace que la red neuronal sea difícil de explicar en su funcionamiento gráfico. Se componen de tres capas principales, capa de entrada, capa oculta y capa de salida (Figura I.4.).

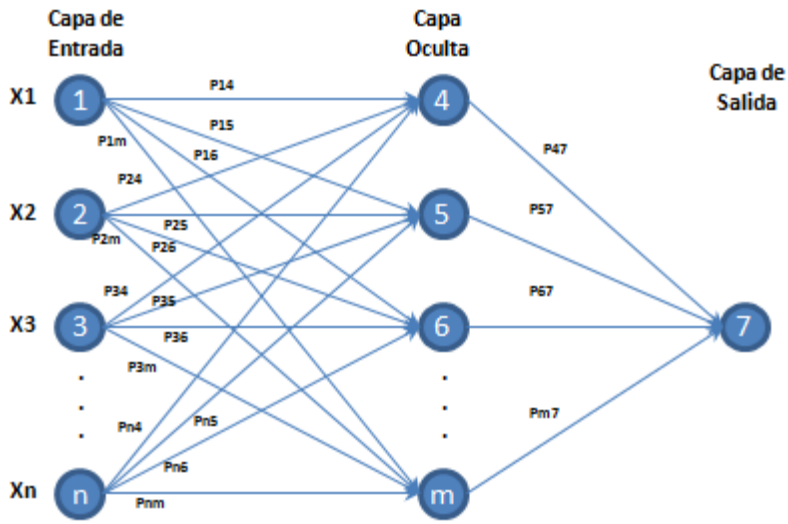


Figura I.4. Ejemplo red neuronal, capas principales.

Mediante conexiones entre los nodos de la capa de entrada (1 … n) se propaga hacia adelante generando 1 o más capas ocultas dependiendo de la complejidad que generen los datos, cada nodo es independiente y al final genera un resultado en el nodo de la capa de salida.

Las conexiones tienen un peso determinado por el algoritmo, P14 significa el peso que existe en la conexión del nodo 1 al nodo 4, esto entre la capa de entrada y la oculta, P47 es el peso de la conexión entre el nodo 4 y el 7. Mediante lo que se llama una función de activación de cada nodo se activa la conexión entre cada uno hasta alcanzar la capa de salida.

Las redes neuronales son de dos tipos[2]:

Aprendizaje supervisado: Utiliza variables de entrada y la objetivo, las primeras entran a la red por la primer capa y hasta llegar a la de salida comparando al final los resultados con la variable objetivo que tiene los resultados reales, los pesos de las capas se ajustan para asegurar una respuesta correcta para casos similares a las variables de entrada.

Existen las redes con un perceptrón simple, quiere decir que no generan capas ocultas, solo la de entrada y la de salida, son equivalentes a una función de discriminante lineal.

Cuando no es posible separar linealmente los conjuntos de datos se utilizan los modelos de Perceptrón Multicapa las cuales tienen una o más capas ocultas.

La forma en la cual trabaja cada nodo o neurona se basa en tres funciones (Figura I.5.)

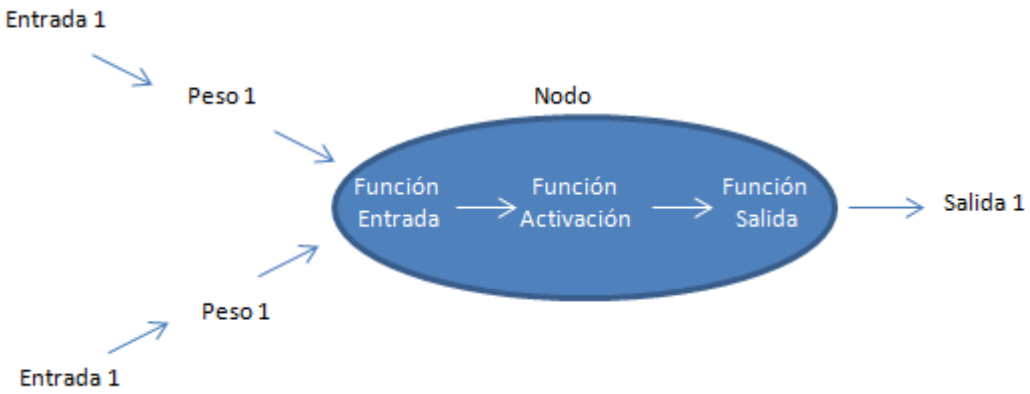


Figura I.5. Ejemplo funciones del nodo o neurona.

[2] José Hernández Orallo, M.José Ramírez Quintana, César Ferri Ramírez Introducción a la Minería de Datos.

La función de entrada, de activación y de salida

La función de entrada recibe los valores con los diferentes pesos asignados. Las más utilizadas son:

- Suma de las entradas pesadas: suma los valores multiplicados por sus pesos

$$\sum (\text{Entrada 1} * \text{Peso 1})$$

- Multiplicación de las entradas pesadas: multiplicación de los valores multiplicados por los pesos.

$$\Pi (\text{Entrada 1} * \text{Peso 1})$$

- Máximo de las entradas pesadas: Considera el valor máximo de la entrada por el peso.

$$\text{Max } (\text{Entrada 1} * \text{Peso 1})$$

La función de activación indica los diferentes estados en los que puede estar un nodo (neurona), el rango va normalmente de 0 a 1 o de -1 a 1, indicando que el nodo está inactivo = [0, -1] ó activo = 1. Las funciones de activación más comunes son:

- Función lineal: $f(x) = \begin{cases} -1 & x \leq -1/a \\ a * x & -\frac{1}{a} < x < 1/a \\ 1 & x \geq -1/a \end{cases}$

  Donde x = función de entrada global – un umbral y a>0

- Función sigmoidea: $f(X) = \frac{1}{1+e^{-gx}}$

  Donde x = función de entrada global – un umbral

- Función tangente hiperbólica: $f(x) = \frac{e^{gx}-e^{-gx}}{e^{gx}+e^{-gx}}$

Donde x = función de entrada global – un umbral

La función de salida es el valor que se le transmitirá a los siguientes nodos, las más comunes son:

- Función identidad. Donde la salida es igual que la entrada
- Binaria: $\begin{cases} 1 & si\ la\ activación\ es\ \geq al\ umbral \\ 0 & de\ lo\ contrario \end{cases}$

Aprendizaje no supervisado: Se tienen solo los datos de entrada sin una variable objetivo, la red debe auto entrenarse dependiendo de estructuras existentes en los datos. Se usan principalmente para reducir la dimensión de los datos y para la agrupación. Para el presente trabajo solo se utilizará la red neuronal como aprendizaje supervisado por lo que no se tratará más a fondo este tipo de red.

Consideraciones sobre estos modelos.

Las redes neuronales son una parte muy importante dentro de la inteligencia artificial, se ha generado mucha teoría al respecto y tiene un gran soporte en la investigación.

El principal problema que enfrenta es su difícil (a veces casi imposible) explicaición en su estructura. Para esto se han desarrollado varias técnicas como lo es explicar su construcción a partir de árboles de decisión para hacerlo más inteligible. Otro problema que tienen este tipo de algoritmos es su lentitud y consumo de recursos computacionales entre más complejos se vuelven.

Cuando se compara en su precisión contra otros modelos generalmente no aparece entre los mejores, sin embargo siempre es bueno considerarlo dentro de los modelos a competir por su robustez teórica.

I.2.3. Redes bayesianas.

Las redes bayesianas se generan a partir del teorema de Bayes (cálculo de una probabilidad condicional en eventos dependientes)[3], con este modelo se puede hacer inferencia, estimar la probabilidad posterior de las variables desconocidas en base a las conocidas, también se conoce como la relación causa-efecto entre las variables.

En su representación gráfica los nodos son variables aleatorias y las conexiones relaciones de dependencia directa entre ellas. En forma gráfica puede parecer un árbol tradicional aunque también puede ser lo que se conoce como un poliárbol (un nodo que puede tener más de un padre).

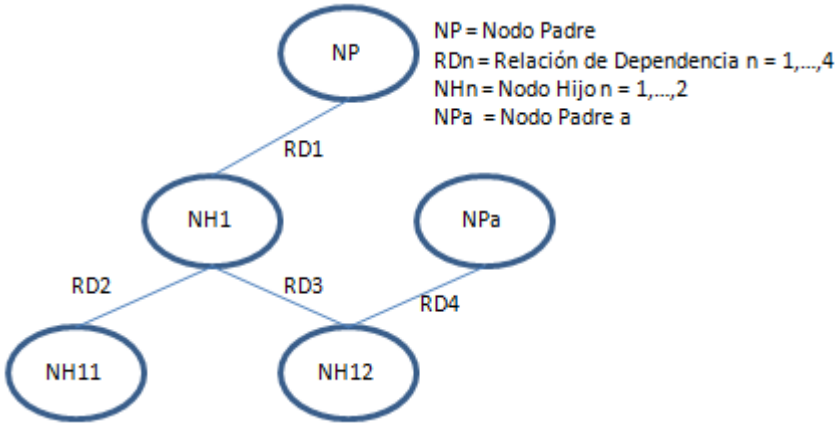


Figura I.6. Ejemplo representación gráfica red bayesiana.

Como se puede observar se puede tener más de un antecedente para la misma consecuencia (NH12 depende de NH1 o NPa), un ejemplo de este tipo de dependencia podría ser una enfermedad, donde un síntoma sea fiebre (NH12) y éste pueda venir por un dolor de estómago (NH1) o por una infección en la garganta (NPa), es decir, dado que infección de garganta entonces fiebre o dado que dolor de estómago entonces fiebre.

Al igual que las redes neuronales también pueden utilizarse para la predicción y para el agrupamiento. Una de sus grandes ventajas con respecto a otras técnicas es que permite calcular de forma explícita la probabilidad asociada a cada una de las hipótesis posibles, es decir cuantifica la incertidumbre.

[3] José Hernández Orallo, M.José Ramírez Quintana, César Ferri Ramírez Introducción a la Minería de Datos.

Las redes bayesianas están basados en el teorema de Bayes que se representa:

$$P(h|O) = \frac{P(O|h)*P(h)}{P(O)} \quad , P(O) \neq 0$$

Donde h es la hipótesis, O es la observación, por lo que P(h|O) es la probabilidad de la hipótesis dado lo observado (probabilidad condicional). Se asume independencia entre las observaciones.

La hipótesis puede tener más de un valor por lo que la red bayesiana busca la que tenga la máxima probabilidad de los valores dadas las observaciones, a esto se le llama la hipótesis máxima a posteriori (máximum a posteriori = MAP).

En términos de un modelo de clasificación se tiene:

- La variable objetivo (VO) o hipótesis (h) con diferentes valores a predecir del cual se tomará el MAP.

- La(s) variable(s) explicativa(s) (VE) o las observación(es) (Oi) donde i es la VE iésima.

La fórmula se expresaría en los siguientes términos:

$$P(h_{MAP}|O_1 \dots O_n) = \frac{P(O_1 \dots O_n|h_{MAP})\ P(h_{MAP})}{P(O_1 \dots O_n)}$$

Tomando el siguiente ejemplo:

Edad	**Ingresos**	**Antigüedad**	**Morosidad**
Mayor	Altos	Antiguo	No
Joven	Bajos	Nuevo	Si
Medio	Altos	Antiguo	No
Joven	Bajos	Nuevo	Si

Mayor	Medios	Antiguo	Si
Joven	Bajos	Nuevo	Si
Medio	Medios	Nuevo	Si
Medio	Medios	Nuevo	No
Medio	Altos	Antiguo	Si
Mayor	Bajos	Nuevo	No

Tabla I.4. Ejemplo Datos para Red bayesiana.

Se tiene un caso nuevo a predecir si será moroso:

Edad	**Ingresos**	**Antigüedad**	**Morosidad**
Joven	Altos	Antiguo	¿?

Tabla I.5. Ejemplo Datos Red bayesiana para predicción de morosidad.

VE = h = Morosidad: Si ó No, la probabilidad para

$$P(Morosidad = Si) = \frac{6}{10} = .6$$

$$P(Morosidad = No) = \frac{4}{10} = .4$$

Por lo que la máxima probabilidad a priori es “Si”. Entonces dado el nuevo caso se reduce a

$$P(h_{MAP}) = P(Si) = .6$$

$$P(Si|Edad, Ingresos, Antiguedad) = \frac{P(Edad|Si)P(Ingresos|Si)P(Antiguedad|Si)P(Si)}{P(Edad, Ingresos, Antiguedad)}$$

Sustituyendo los valores de las variables nuevas se tiene:

$$P(Si|Joven, Altos, Antiguo) = \frac{P(Joven|Si)P(Altos|Si)P(Antiguo|Si)P(Si)}{P(Joven, Altos, Antiguo)}$$

$$P(Joven|Si) = \frac{3}{6} = .5$$

$$P(Altos|Si) = \frac{1}{6} = .16666667$$

$$P(Antiguo|Si) = \frac{2}{6} = .3333333$$

$$P(Joven, Altos, Antiguo) = 1\,, probabilidad\ trivial$$

Sustituyendo valores se tiene

$$P(Si|Joven, Altos, Antiguo) = \frac{.5 * .16666667 * .3333333 * .6}{1} = .0177$$

La probabilidad de ser moroso el nuevo caso es de .0177, de igual forma se puede calcular la probabilidad de Morosidad="No".

Consideraciones sobre estos modelos.

Este tipo de técnicas tiene una profunda metodología teórica que lo hace imprescindible dentro de las opciones a considerar.

En el capítulo IV se incluirán estos modelos para mostrar sus resultados y compararlo con otros algoritmos aquí mencionados.

I.2.4. Técnicas estadísticas.

Las técnicas estadísticas clásicas para la predicción tienen ya muchos años, son ampliamente conocidos por sus buenos resultados. Se revisarán las técnicas paramétricas y no paramétricas.

I.2.4.1. Técnicas paramétricas.

Estas técnicas se utilizan en muchos casos para la predicción, al igual que las técnicas antes vistas se tiene una variable objetivo o dependiente denotada por Y, las variables de entrada o independientes representadas generalmente por la letra X.

Modelos de regresión.

El modelo más simple es el de ***regresión lineal***, el cual se representa por la siguiente fórmula:

$$y_i = C_0 + C_1 x_{i1} + \cdots + C_p x_{ip} + E_i$$

Donde: y_i es la variable dependiente del caso i

$C_0, C_1 \ldots, C_p$ son las constantes que se multiplicaran por las variables independientes para predecir el valor de y_i

$x_{i1}, \ldots, x_{ip}$ son las variables independientes

E_i es el error de la regresión

Viéndolo de manera gráfica para un problema de una variable dependiente Y y una independiente X se tiene

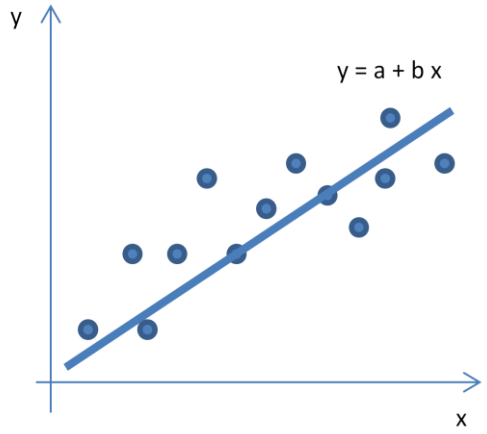


Figura I.7. Ejemplo representación gráfica regresión lineal.

Los puntos son los datos graficados en la coordenadas de la intersección entre las variables, la recta es la regresión lineal que mejor se ajusta a los datos por lo que será la ecuación del modelo para predecir futuros datos.

Se asume linealidad entre los datos de entrada y salida además también se necesitan valores numéricos tanto para variable dependiente e independiente, esto podría ser un problema porque en general las variables independientes no siempre son lineales, para estos casos se pueden utilizar diversas técnicas de transformación de datos.

El problema es encontrar la recta que minimice el error residual o distancias o desviaciones entre la recta y los puntos. El error está definido por la resta de la distancia entre el punto real y el que aproxima el modelo, esto se conoce como el error cuadrático medio

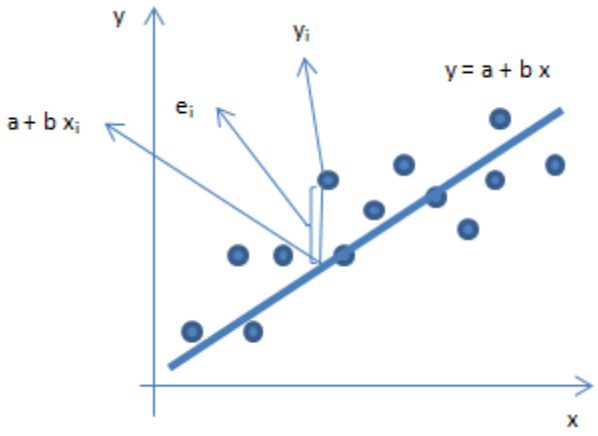


Figura I.8. Ejemplo regresión lineal distancia entre los puntos y la recta.

$$\sum e_i^2 = \sum [y_i - (a + bx_i)]^2$$

Donde: e_i^2 es el error cuadrático medio para el i ésimo punto

y_i es el punto i de los datos originales

$a + bx_i$ es la función lineal para el punto i

Lo que se debe minimizar es $\min (\sum e_i^2)$

Esto representa la variabilidad que no es explicada por el modelo

Modelos lineales generalizados (MLG).

Muchos de los problemas a enfrentar para algoritmos predictivos no cumplen con los supuestos lineales de la regresión como que la variable objetivo es numérica o que su relación con las variables explicativas se podría explicar mediante una función lineal, o que sigue una distribución normal.

Los modelos lineales generalizados (Figura I.9.) permiten modelar situaciones sin los supuestos antes mencionados. Los supuestos para esos modelos son que la variable dependiente o de respuesta sigue una distribución exponencial así mismo el error y se tiene una función que relaciona con las variables explicativas la cual generalmente es no lineal.

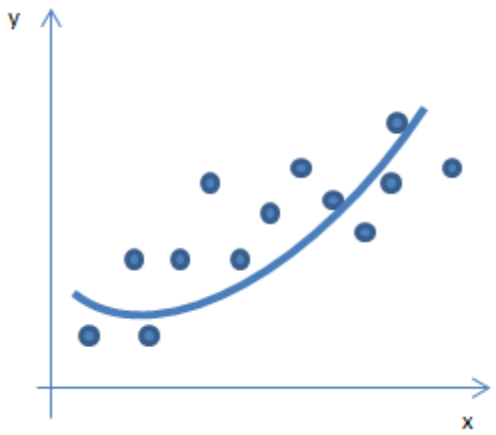


Figura I.9. Ejemplo modelos lineales generalizados.

$$y_i = r(C_0 + C_1 x_{i1} + \cdots + C_p x_{ip}) + \varepsilon_i$$

$$r\left(C_0 + C_1 x_{i1} + \cdots + C_p x_{ip}\right) = E\left[y_i \middle| x_{i1} \ldots x_{ip}\right] = \mu_i\,, \qquad E[\varepsilon_i] = 0$$

La distribución es exponencial si su función de densidad o probabilidad es como sigue

$$f(y, \beta) = Exp\,[a(y)b(\beta) + c(\beta) + d(y)]$$

Distribuciones que cumplen estos requisitos son: la normal, binomial, Poisson y gamma.

La ***regresión logística*** es uno de los modelos más populares al utilizarse dentro de los MLG ya que genera una probabilidad para la variable de salida, donde la suma de todas será igual a uno.

Para el caso más sencillo se puede tener una variable dependiente binaria de la siguiente forma:

$$y_i = \begin{cases} 1, & Prob_i(1) = p_i \\ 0, & Prob_i(0) = 1 - p_i \end{cases}$$

La variable dependiente tiene dos valores con parámetros 1 y p_i que es la probabilidad de 1

El valor esperado del modelo es:

$$\mu_i = E[y_i] = [1 \times p_i] + [0 \times (1 - p_i)] = \ p_i$$

La función logística es la que se convertirá en la unión utilizando una transformación lineal del predictor línea para [1, 0]:

$$log\,\frac{p_i}{1 - p_i}$$

El ***análisis discriminante*** es otra técnica dentro de los MLG que busca encontrar reglas de asociación entre objetos a una clase ya establecida, es una técnica estadística para la clasificación supervisada.

A partir de q grupos donde se asignan a una serie de objetos y de p variables medidas sobre ellos $(x_1, \dots, x_p)$, se trata de obtener para cada objeto una serie de puntuaciones que indican el grupo al que pertenecen $(y_1, \dots, y_m)$, de modo que sean funciones lineales de $x_1, \dots, x_p$

$$y_i = a_{i1}x_1 + \cdots a_{ip}x_p + a_{i0}$$

$$\dots$$

$$y_m = a_{m1}x_1 + \cdots a_{mp}x_p + a_{m0}$$

Donde m = mín(q-1, p) tales que discriminen al máximo los q grupos. Las combinaciones lineales de las p variables deben maximizar la varianza entre los grupos y minimizarla dentro de los grupos.

La $corr(y_i, y_m) = 0$ para todo $i \neq m$

Cuando las variables $x_1, \dots, x_p$ están tipificadas entonces las funciones $y_i = a_{i1}x_1 + \cdots a_{ip}x_p$ para i = 1 … m, se denominan funciones discriminantes canónicas.

Las funciones $y_1, \dots, y_m$ se extraen de modo que:

- y_1 sea la combinación lineal de $x_1, \dots, x_p$ que proporciona la mayor discriminación posible entre los grupos.

- y_2 sea la combinación lineal de $x_1, \dots, x_p$ que proporciona la mayor discriminación posible entre los grupos después de $y_1$tal que $corr(y_i, y_m) = 0$

En general y_i es la combinación lineal de $x_1, \dots, x_p$ que proporciona la mayor discriminación posible entre los grupos después de y_{i-1} y tal que $corr(y_i, y_j) = 0$ para j = 1,…, (i - 1)-

I.2.4.2. Técnicas no paramétricas.

Las técnicas no paramétricas generan modelos que pueden ajustar mejor a los datos debido a que son más flexibles que las técnicas paramétricas permitiendo resaltar los patrones de dependencia presentes en los datos.

Para un modelo de regresión simple se tiene

$$y_i = r(x_i) + \epsilon_i$$

y_i es la variable dependiente continua con solo hay una variable independiente x_i, $r(x_i)$ es la función de regresión que se supone regular

$$E(\epsilon_i) = 0 \; y \; V(\epsilon_i) = \sigma^2 \;\; para \; todo \; i$$

Donde el E = error y V = varianza, la cual es constante.

La hipótesis de los errores de la varianza constantes pueden suponer que está en función de la variable independiente x.

El siguiente paso es estimarlo para las n observaciones disponibles, construyendo un estimador $\check{r}(x)$ de la función y de la varianza del error σ^2.

Algunas de las técnicas para estimar la función de regresión son:

- Ajuste local de modelos paramétricos. Realizan varios ajustes paramétricos tomando en cuenta solamente los datos cercanos al punto de estimación de la función.

- Métodos basados en series ortogonales de funciones. Se elige una base otonormal del espacio vectorial de funciones y se estiman los coeficientes del desarrollo en esa base de la función de regresión.

- Suavizado mediante splines. Se busca la estimación de la función que minimiza la suma de los cuadrados de los errores más un término que penaliza la falta de suavidad.

- Técnicas de aprendizaje supervisado. Pueden utilizarse redes neuronales, k vecinos más cercanos y árboles de regresión para estimar r(x).

Estimadores de núcleo y ajuste local de polinomios.

Se busca generar regresiones lineales en intervalos que se ajusten mejor a los datos para obtener una serie de líneas rectas.

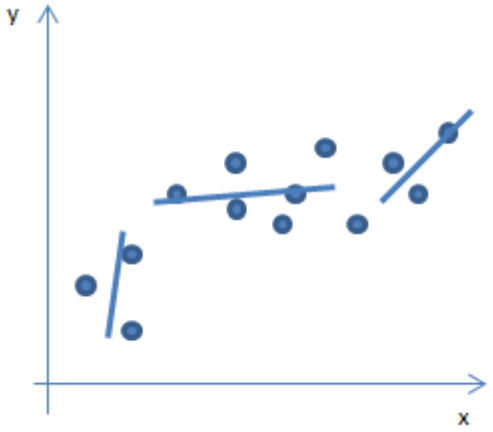


Figura I.10. Ejemplo gráfica de estimadores de núcleo y ajuste local de polinomios.

La relación entre las variables es aproximadamente lineal en los intervalos, la regresión estimada es discontinua, para que esto no sucede se deben de tomar intervalos de la variable independiente muy cercanos a los intervalos donde es discontinua y así tener un ajuste lineal local.

Figura I.11. Ejemplo gráfica de estimadores de núcleo y ajuste local de polinomios no completamente continuas.

Pese a que las regresiones ya se acercan mucho aún no son continuas para sus valores extremos.

Esto puede conseguirse ponderando los datos de forma que el peso w de una observación (x, y) sea función decreciente de la distancia de su ordenada x al punto t donde se realiza la estimación.

Usualmente se utiliza una función núcleo (kernel) K, el peso de la estimación será:

$$w_i = w(t, x_i) = \frac{K\left\{\frac{x_i - t}{h}\right\}}{\sum_j^n K\left\{\frac{x_j - t}{h}\right\}}$$

Donde h es un parámetro de escala que controla la concentración del peso total alrededor de t, a h se le denomina parámetro de suavizado. Una vez determinados los pesos se debe de resolver el problema de mínimos cuadrados siguiente:

$$\min(a, b) \sum_{i=1}^{n} w_i (y_i - (a + b(x_i - t)))^2$$

La recta de regresión ajustada localmente alrededor de t es:

$$l_t(x) = a(t) + b(t)(X - t)$$

La estimación de la función de regresión en el punto t es el valor que toma esa recta en x=t

$$\check{r}(t) = l_t(t) = a(t)$$

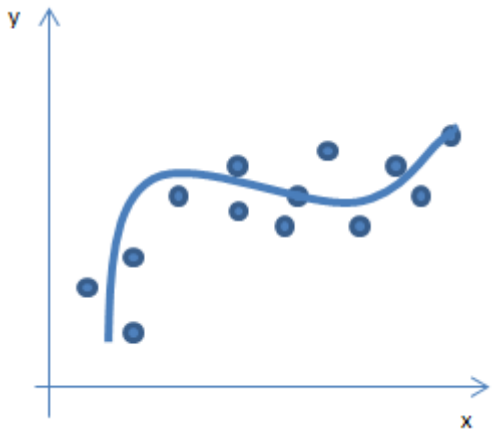


Figura I.12. Ejemplo gráfica de estimadores de núcleo y ajuste local de polinomios suavizado.

Las funciones de núcleo más usadas son:

Uniforme: $\frac{1}{2} I_{[-1,1]}(x)$

Gaussiano: $\left(\frac{1}{\sqrt{2\pi}}\right) exp\left(\frac{-x^2}{2}\right)$

Epanechnikov: $\left(\frac{3}{4}\right)(1-x^2) I_{[-1,1]}(x)$

Biweight: $\left(\frac{15}{16}\right)(1-x^2)^2 I_{[-1,1]}(x)$

Donde la función $I_{[a,b]}$ es la función unidad (valor 1) en el intervalo (a, b).

El estimador lineal local se puede generalizar al ajuste local de regresiones polinomiales de mayor grado q. De una variable x se obtiene el polinomio:

$$\beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_q x^q$$

Y se actúa como en una regresión lineal con q regresores.

Elección del parámetro de suavizado.

La elección del parámetro de suavizado h tiene una gran importancia en el aspecto y propiedades del estimador de la función de regresión, controla el equilibrio que la función de regresión debe mantener entre el buen ajuste a los datos observados y la capacidad de predecir bien observaciones futuras.

Puede elegirse de forma manual comparando los resultados obtenidos para distintos valores de h y eligiendo el que da el mejor resultado visual, lo cual es subjetivo, para la elección están los siguientes métodos más habituales:

- Error de predicción en una muestra de prueba. Si se tiene una muestra suficiente de datos de entrenamiento y de comprobación se pude usar el error cuadrático medio:

$$ECMP_{prueba}(h) = \frac{1}{n_T} \sum_{i=1}^{n_{prueba}} (y_i^{prueba} - \hat{r}_h(x_i^{prueba}))^2$$

Donde $(y_i^{prueba}, x_i^{prueba})$, $i = 1, \dots, n_{prueba}$ es la muestra de prueba y $\hat{r}_h$ es el estimador no paramétrico construido con parámetro suavizado h usando la otra parte de la muestra original. Debe elegirse el parámetro suavizado h_{prueba} que minimice la función.

- Validación cruzada. Técnica usada para la elección de parámetros que controlan la bondad de ajuste y la capacidad predictiva cuando no hay muestra de prueba. Se utiliza una muestra del total de los datos como prueba y se valida contra la real, así sucesivamente para otra muestra hasta que se minimice el error.

$$ECMP_{VC}(h) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{r}_h(x_i))^2$$

Mide el error de estimación del predictor fuera de la muestra para cada h. El valor que minimice esa función h_{CV} como el parámetro de suavizado.

- Validación cruzada generalizada. Modificación a la validación cruzada que se aplica a los estimadores que son lineales a los casos de la variable explicada.

$$ECMP_{GVC}(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{y}_i}{\frac{1 - Traza(S)}{n}}\right\}^2$$

Donde $Traza(S) = \sum_{i=1}^{n} S_{ii}$ es la suma de los elementos de la diagonal S.

Regresión múltiple. Modelos Aditivos.

Estas regresiones se ocupan cuando hay múltiples variables independientes, en el caso de p variables independientes se tiene:

$$y_i = r(x_{i1}, \dots, x_{ip}) + \epsilon_i\,, \;\; con\; E(\varepsilon_i) = 0 \;\; y \;\; V(\varepsilon_i) = \sigma^2 \;\; para\; todo\; i = 1, \dots, n$$

Ahora la función r será polinomial, para estimarla se necesita calcular los pesos w_i de cada observación para cada variable explicativa del modelo local.

Para estimar r en el punto $t = (t_1, \dots, t_d)$ para los datos $(y_i; x_{i1}, \dots x_{ip})$ los que más peso tienen con las variables independientes $x = (x_1, \dots x_p)$ más cercanos a $t = (t_{1,} \dots t_p)$, como es polinomial las distancias entre x y t se debe medir entre un espacio de dimensión p.

Una forma de asignar pesos w_i es:

$$w_i = (t, x_i)\; \alpha \prod_{j=1}^{p} K\left[\frac{x_{ij} - t_j}{h_j}\right]$$

Donde K es un núcleo univariante, h_j es el parámetro de suavizado adecuado para la j-ésima variable dependiente y α indica proporcionalidad.

Para una alta dimensionalidad, es decir muchas variables independientes, se puede usar un modelo aditivo de la forma:

$$y_i = \alpha + \sum_{j=1}^{p} g_i(x_{ij}) + \epsilon_i$$

Con $E(\varepsilon_i) = 0$ y $V(\varepsilon_i) = \sigma^2$ para todo $i = 1, \dots, n$ y $E\left(g_j(x_j)\right) = 0$ con $j = 1, \dots, p$

Regresión binaria local.

Para este tipo de regresión la variable dependiente será categórica binaria, se le conoce también como la versión no paramétrica de los modelos lineales no generalizados.

Por ejemplo, para dos clases $C_0\ y\ C_1$ con los casos $(y_i; x_{i1}, \dots x_{ip})$ con $i = 1, \dots, n$

Los valores para $y = 1\ o\ 0$ representada por una variable aleatoria con probabilidades de p y 1-p respectivamente. Donde p es la función de las variables independientes $(x_{ij}, \dots, x_{ip})$:

$$E(y_i) = P(y_i = 1) = p_i = r(x_{ij}, \dots, x_{ip})$$

y_i sigue una distribución Bernoulli de parámetro $p_i = r(x_{ij}, \dots, x_{ip})$

Si $x_i = (x_{i1}, \dots, x_{ip})$ está en un entorno $x = (x_1, \dots, x_p)$ entonces y_i sigue el siguiente modelo logístico:

$$p_i = \frac{1}{1 + e^{-\beta^t x_i}}\ ,\quad \text{ó}\quad log\left[\frac{p_i}{1 - p_i}\right] = \beta^t x_i$$

Se puede estimar en un modelo paramétrico de máxima verosimilitud, la contribución de cada observación a la función de log-verosimilitud es:

$$y_i log\left[\frac{p_i}{1 - p_i}\right] + \log(1 - p_i)$$

Sumando las observaciones y ponderando por un peso $w_i = w(x,\ x_i)$ se obtiene la función de log verosimilitud local:

$$l_x(\beta) = \sum_{i=1}^{n} w_i \left[y_i log\left[\frac{p_i}{1 - p_i}\right] + \log(1 - p_i)\right]$$

El estimador de $\beta, \hat{\beta} = \hat{\beta}(x)$ se obtiene maximizando la función anterior y permite obtener una estimación de $p_x = r(x_1, \dots, x_p)$:

$$\hat{r} = (x_1, \dots, x_p) = \widehat{p_x} = \frac{1}{1 + e^{-\beta^t x}}$$

Si es mayor o menor a 0.5 se clasificará la observación en $x = (x_1, \dots, x_p)$ en C_0 o en C_1 respectivamente.

Consideraciones sobre estos modelos.

Los modelos clásicos de estadística son ampliamente conocidos, con el profundo sustento teórico matemático que los hace imprescindibles dentro de cualquier consideración de modelos predictivos a evaluar.

Desde su uso para encontrar las variables significativas para otros modelos como para su elección final como el ganador para ser usado en el problema a resolver, son muy importantes y son el sustento de nuevas técnicas como la minería de datos.

El objetivo del trabajo se centra en las técnicas de minería de datos por lo que se utilizarán en diferentes partes del capítulo III y IV como un complemento a comparar.

I.3. Aprendizaje no supervisado.

Busca describir los datos, comportamientos en donde se puede agrupar, clasificar generar asociaciones entre la información a partir de patrones encontrados con técnicas específicas, no existe una variable objetivo a predecir, sino que busca saber si un conjunto de datos con ciertas características se encuentra en un grupo o en otro o si la secuencia de eventos dentro de los datos seguirá una tendencia similar en otros.

El aprendizaje no supervisado[4] genera clases o grupos en los datos que ayuda a encontrar ciertos patrones de comportamiento desconocidos y útiles, existen dos tipos de técnicas: la asociación y la agrupación:

A diferencia del aprendizaje supervisado no existe una variable objetivo a pronosticar, es decir, no generará un modelo predictivo.

I.3.1. Reglas de asociación.

Relacionan una conclusión con un conjunto de condiciones, su objetivo es encontrar correlaciones entre variables, patrones entre los datos que permitan conocer el comportamiento general del problema a resolver, con ayuda de estos resultados se puedan tomar decisiones.

Una implicación de la forma A entonces B, un ejemplo clásico de asociación se da en una canasta de mercado:

¿Cuál es el comportamiento de un cliente que compra un producto A y también compra un producto B?

Por ejemplo si compra pan entonces comprará también mayonesa, si se tienen suficientes eventos en donde los clientes eligen productos y siempre llevan los dos antes mencionados entonces el algoritmo encontrará una regla donde A implica B, es decir hay una fuerte asociación entre estos dos productos.

[4] José Hernández Orallo, M.José Ramírez Quintana, César Ferri Ramírez Introducción a la Minería de Datos.

Para el siguiente ejemplo ficticio se muestran las siguientes compras de 5 clientes en un supermercado.

Cliente	Compra
C1	Pan, carne, jitomate
C2	Pan, mayonesa, jamón, huevo, crema
C3	Pan, queso, leche
C4	Pan, mayonesa, crema, mermelada
C5	Pan, queso, leche, mayonesa, aguacate

Tabla I.6. Ejemplo compras de 5 clientes en un supermercado.

Para analizar las asociaciones entre los productos se debe armar una tabla con la información de los clientes en cada fila y en las columnas cada uno de los posibles productos a comprar representados por una variable binaria que indica si el producto fue adquirido por el cliente, por ejemplo:

Cliente	Pan	Mayonesa	Carne	Jitomate	Queso	Jamón	Huevo	Leche	Crema	Mermelada	Aguacate
C1	1	0	1	1	0	0	0	0	0	0	0
C2	1	1	0	0	0	1	1	0	1	0	0
C3	1	0	0	0	1	0	0	1	0	0	0
C4	1	1	0	0	0	0	0	0	1	1	0
C5	1	1	0	0	1	0	0	1	0	0	1

Tabla I.7. Ejemplo datos para analizar las asociaciones entre productos y clientes.

El algoritmo de asociación podría encontrar una relación fuerte entre Pan, Mayonesa y Crema.

La regla de asociación generada por el algoritmo podría ser: si $\propto => \beta$, si Pan y Mayonesa entonces Crema, en donde $\propto y \beta$ son elementos disjuntos.

La calidad de las reglas generadas se miden por:

Soporte: Número de casos o porcentaje de los mismos que la regla predice correctamente.

Precisión: Mide el porcentaje de veces que la regla se puede aplicar a casos nuevos.

Para el ejemplo anterior se tendría un soporte igual a 3 y una precisión del 67%.

Las reglas de asociación no soportan el uso de variables numéricas solo trabajan con categóricas, para poder sortear esta dificultad se pueden realizar conversiones (discretizar) los valores numéricos.

Algunos algoritmos de este tipo son:

Inducción de reglas generalizado (GRI por sus siglas en inglés): Extrae reglas con el mayor nivel de contenido de información basándose en un índice que tiene en cuenta tanto la generalidad (soporte) como la precisión (confianza) de las reglas.

A priori: Destaca las reglas con un mayor contenido de información, utiliza un esquema de indización para grandes volúmenes de información.

Secuencia: Encuentra reglas de asociación en datos secuenciales u ordenados en el tiempo por lo que tiene un orden previsible.

I.3.2. Modelos de agrupación o clusters.

Las técnicas de agrupación buscan identificar grupos similares y etiquetan los casos en base al grupo donde pertenecen, se basan en la medición de las distancias entre registros y entre grupos.

No hay variable dependiente u objetivo para el modelo a pronosticar, por lo que se genera si n tener conocimientos previos sobre las características de los grupo, además de no saber cuántos grupos se deben buscar.

Los casos se asignan a los grupos de un modo que tiende a minimizar la distancia entre los casos pertenecientes al mismo grupo, éstos con frecuencia pueden ser usados como insumos en otro tipo de modelos (Figura I.13.).

Su utilidad consta en encontrar agrupaciones interesantes en los datos que proporcionen descripciones útiles para la toma de decisiones.

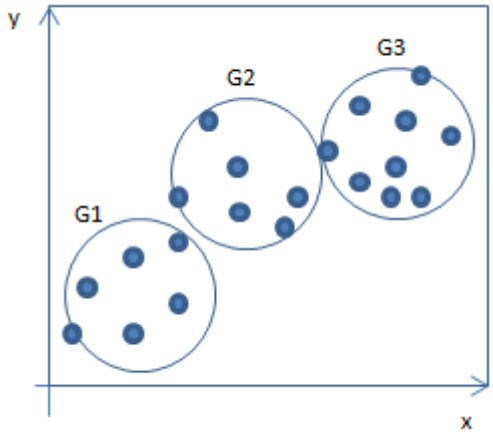


Figura I.13. Ejemplo agrupación de casos.

Se encuentran 3 grupos para los datos de ejemplo ficticios los cuales se separan por el centro de cada uno de los círculos que delimitan los grupos.

Para encontrar la similitud entre los grupos es necesario encontrar una medida de distancia para calcular ésta entre los casos de los grupos.

Las medidas de distancia tradicionales que aplican sobre casos numéricos son:

- Distancia Euclídea. Longitud de la recta que une dos puntos en el espacio euclídeo.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- Distancia de Manhattan. El llegar del punto x al y no en diagonal sino zizagueando.

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i|$$

- Distancia de Chebychev. Divergencia más grande en alguna de las dimensiones.

$$d(x,y) = max_{i=1\ldots n}|x_i - y_i|$$

- Distancia del coseno. El coseno del ángulo que forma considerando cada caso como un vector.

$$d(x,y) = arccos\left[\frac{x^T \times y}{||x|| \times ||y||}\right]$$

- Distancia de Mahalanobis. Utiliza la matriz de covarianzas S.

$$d(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$$

Para todas las medidas de distancia anteriores, lo mejor es normalizar todos los atributos de los casos para que pesen lo mismo y no generen distancias mucho mayores que otras.

Con datos no numéricos también se puede utilizar un concepto de distancia utilizando una función delta $\delta(a,b) = 0$ si y solo si $a = b$ y $\delta(a,b) = 1$, utilizando la siguiente función:

$$d(x,y) = \omega \sum_{i=1}^{n} \delta(x_i, y_i)$$

Donde ω es un factor de reducción.

Si los datos son conjuntos se puede definir la distancia entre éstos por:

$$d(x,y) = \frac{|x \cup y| - |x \cap y|}{|x \cup y|}$$

Algunos algoritmos de este tipo:

K-medias:

Es un método de agrupamiento por vecindad. Selecciona conjuntos de datos en grupos distintos. El método define un número fijo de grupos, de forma iterativa asigna registros a los grupos y ajusta sus centros hasta que no se pueda mejorar el modelo.

Puede seguir dos enfoques:

Por lotes: Cuando todos los datos de entrada están disponibles desde el principio.

En línea: No se dispone de todos los datos al principio y se añadirán más casos después.

En su fase de entrenamiento se sigue el siguiente procedimiento:

1. Se calcula x_k el prototipo más próximo a A_g y se incluye en la lista de ejemplos del prototipo: $A_g = arg_{A_i} \min[d(x_k, A_i)]\,,\ \forall i = 1 \dots n$

2. Después de haber introducido todos los casos, cada ejemplo A_k tendrá un conjunto de ejemplos a los que representa: $l(A_k) = [x_{k1}, x_{k2}, \dots x_{kn}]$

3. Se desplaza el prototipo hacia el centro de masas de su conjunto de ejemplos: $A_k = \frac{\sum_{i=1}^{m} x_{k_i}}{m}$

4. Se repite el proceso hasta que ya nos se desplacen los prototipos.

Jerárquico:

Se basan en la construcción de un árbol en el que las hojas con los elementos del conjunto de casos y el resto son subconjuntos de ejemplos que pueden ser utilizados como particionamiento del espacio (Figura I.14.). También denominado dendrograma.

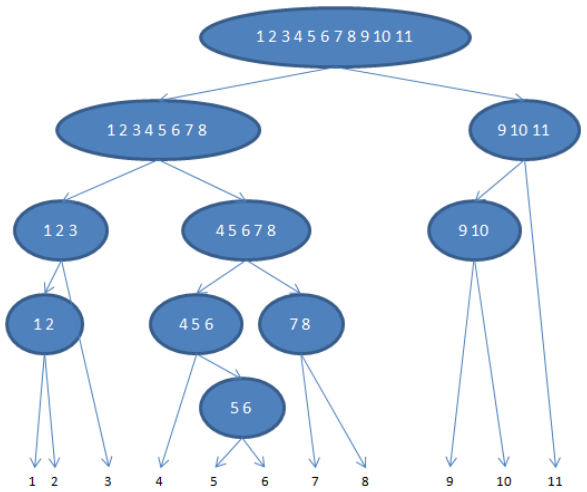


Figura I.14. Ejemplo agrupación modelo jerárquico.

Los niveles son representados por las hojas, para este ejemplo son 10 niveles.

Los métodos se dividen en:

Aglomerativos: Se construye de abajo hacia arriba, comenzando por las hojas finales y terminando en la raíz.

Divisivos: Se construye de arriba abajo desde la raíz hasta las hojas finales.

El método sigue los siguientes pasos:

1. Hacer que cada punto sea el representante de un grupo que solo contiene dicho punto.
2. Calcular las distancias entre todos los grupos existentes dos a dos.
3. Elegir los dos grupos cuya distancia sea menor.
4. Mezclar los grupos elegidos en el paso anterior. Si el representante de uno de los grupos es el vector: $\overline{C_a} = \{c_{a_1}, c_{a_2}, \dots, c_{a_n}\}$ y del otro grupo $\overline{C_b} = \{c_{b_1}, c_{b_2}, \dots, c_{b_n}\}$, si además el grupo a tiene j casos y el grupo b tiene k casos, el nuevo representante se calculará mediante la expresión:

$$\bar{C} = \left\{\frac{jc_{a_1} + kc_{b_1}}{j+k}, \frac{jc_{a_2} + kc_{b_2}}{j+k}, \dots, \frac{jc_{a_n} + kc_{b_n}}{j+k}\right\}$$

5. Si hay más de un grupo ir a 2.

Kohonen:

Genera un tipo de red neuronal de dos capas que se puede usar para agrupar un conjunto de datos en grupos distintos. Cuando la red se termina de entrenar, los registros que son similares se deberían presentar juntos en el mapa de resultados, mientras que los registros que son diferentes aparecerían aparte.

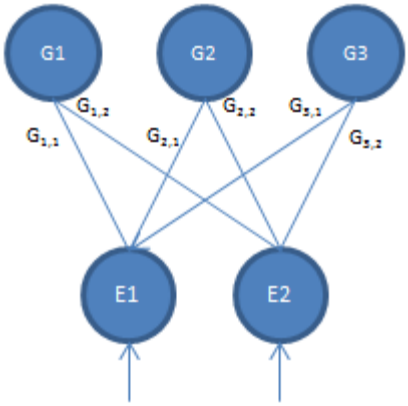


Figura I.15. Ejemplo agrupación modelo Kohonen.

El modelo consta de dos capas, la de entrada donde se introducen los casos y otra de competición en la que cada caso representa un prototipo de agrupación.

Utilizando casos basados en distancia utiliza el entrenamiento definido por un vector de n componentes que puede representarse como un punto en el espacio de n dimensiones, para n=2 sería un punto en n plano.

El procedimiento general es:

1. Se selecciona un ejemplo del conjunto de entrenamiento.
2. Calcular la distancia del caso a cada una de los prototipos.
3. Etiquetar como ganador al prototipo cuya distancia es menor.

Consideraciones sobre estos modelos.

Los modelos de aprendizaje no supervisado tratan de describir los datos, encontrando asociaciones y/o grupos que dependiendo del algoritmo utilizado ayuden en su clasificación.

Al no tener una variable objetivo o dependiente no generan modelos predictivos por lo que el presente trabajo solo los menciona como una parte importante de la minería de datos.

I.4. Recapitulación.

En este capítulo se expuso una breve descripción de lo que es la minería de datos, sus diferentes algoritmos tanto predictivos como descriptivos.

El objetivo fue dar una pequeña introducción a las técnicas y sus fundamentos, no pretende ser de ninguna manera exhaustivo, ya que de cada técnica puede tener libros completos que explican su funcionamiento.

El utilizar la minería de datos se ha vuelto muy popular, esto debido al aumento considerable en la cantidad de información que se almacena actualmente en los repositorios de información de una empresa o institución.

La variedad de algoritmos que manejan, junto a lo amigable que son los software en general para el usuario, hacen de esta disciplina una opción muy viable para proyectos que busquen encontrar información valiosa dentro de sus datos.

Para este trabajo se utilizará software de SPSS, sin embargo hay una gran variedad de herramientas de distintos proveedores como SAS, IBM, Oracle, por mencionar las que se compran o rentan, también existen herramientas de uso libre como lo es WEKA.

Capítulo II. Las devoluciones y la cobranza en el SAT.

En nuestro país la Administración Pública Federal se encarga de la Administración Tributaria , ésta es la encargada de la auditoría, planeación, recaudación y control de los ingresos fiscales; en general, de la aplicación y vigilancia del cumplimiento de las leyes fiscales.

El Servicio de Administración Tributaria (SAT) es un órgano desconcentrado de la Secretaría de Hacienda y Crédito Público (SHCP) que tiene el carácter de Autoridad Fiscal.

Tiene encomendado el despacho de los asuntos, ejecución de las atribuciones y facultades que le confiere la Ley del SAT y su Reglamento Interior, entre las que se encuentran la de Devolución de Impuestos y la Cobranza.

En la Ley del SAT se especifica que debe cumplir los siguientes objetivos.

Artículo 2. “El Servicio de Administración Tributaria tiene la responsabilidad de aplicar la legislación fiscal y aduanera con el fin de que las personas físicas y morales contribuyan proporcional y equitativamente al gasto público, de fiscalizar a los contribuyentes para que cumplan con las disposiciones tributarias y aduaneras, de facilitar e incentivar el cumplimiento voluntario de dichas disposiciones, y de generar y proporcionar la información necesaria para el diseño y la evaluación de la política tributaria”.

La ley fundamental es la Constitución Política de los Estados Unidos Mexicanos, en ella se establecen los derechos y obligaciones de sus habitantes y de sus gobernantes. Se trata de la norma jurídica suprema, ninguna otra ley o disposición puede contrariarla.

Es en nuestra Constitución Política donde se establece la obligación de los mexicanos de contribuir para el gasto público del país, es decir, de pagar impuestos.

Según el precepto constitucional:

“Artículo 31. Son obligación de los mexicanos: … “

“Fracción IV. Contribuir para los gastos públicos, de la Federación como del Estado y Municipio en que residan, de la manera proporcional y equitativa que dispongan las leyes”.

El artículo 2 fracción I del Código Fiscal de la Federación, establece la definición de impuestos:

“Impuestos son las contribuciones establecidas en Ley que deben pagar las personas físicas y morales que se encuentren en la situación jurídica o de hecho prevista por la misma y que sean distintas de las Aportaciones de seguridad social, contribuciones de mejora y Derechos”

La Constitución Política ha previsto que las contribuciones deben observar algunos principios que garanticen que no se haga un cobro indebido de los impuestos. Estos principios forman parte del sistema de valores humanos, tienden a armonizar todas las actividades de los hombres. Dichos principios son los siguientes:

Principio de constitucionalidad.

Uno de los pilares del Derecho Fiscal está constituido por el llamado principio de constitucionalidad, el cual, en términos generales, implica que la relación jurídico tributaria debe encontrarse fundada en los correspondientes preceptos constitucionales, o al menos, debe evitar contradecirlos.

Principio de legalidad.

Los impuestos se fijan y se regulan a través de leyes. El principio de legalidad se refiere a que no se podrá cobrar ningún impuesto o contribución que no se encuentre establecido en una ley, con anterioridad al hecho o circunstancia que fue la causa del pago de un impuesto.

Principio de obligatoriedad.

El principio de obligatoriedad en materia fiscal tiene que entenderse en función no sólo de la existencia de un simple deber a cargo de los contribuyentes, sino como una auténtica obligación pública, de cuyo incumplimiento pueden derivarse severas consecuencias para el beneficio social y económico del país.

Principio de proporcionalidad.

El Principio de Proporcionalidad considera que los contribuyentes deben contribuir a los gastos públicos en virtud de sus capacidades económicas, aportando a la hacienda pública una parte justa y adecuada de sus ingresos, utilidades o rendimientos, pero nunca una cantidad tal, que su contribución represente prácticamente el total de los ingresos netos que haya percibido, pues en este último caso se estaría utilizando a los tributos como un medio para que el Estado confisque los bienes de sus ciudadanos.

Principio de equidad.

El principio de equidad significa la igualdad, ante la misma ley tributaria, de todos los contribuyentes sujetos a un mismo tributo, quienes en tales condiciones, deben recibir un tratamiento idéntico, debiendo únicamente variar las tarifas tributarias aplicables de acuerdo con la capacidad económica de cada contribuyente.

Una vez que una persona física o moral empieza a desempeñar una actividad profesional obteniendo una remuneración económica, tendrá la obligación de inscribirse en el Registro Federal de Contribuyentes (RFC), contrayendo así ciertas obligaciones fiscales, de conformidad con lo señalado en las disposiciones en materia fiscal. Algunas de las obligaciones existentes son:

- En primera instancia, la inscripción ante el Registro Federal de Contribuyentes en el régimen que le corresponda de acuerdo con la actividad que desempeñe
- Llevar contabilidad, de conformidad con el Código Fiscal de la Federación y su Reglamento.
- Expedir y conservar comprobantes con requisitos fiscales, los cuales deberán amparar los ingresos que perciban, así como los gastos que pretendan deducir.
- Determinar los impuestos correspondientes, a través de la presentación de pagos provisionales y declaración anual.
- Presentar las declaraciones informativas que correspondan.

En este tenor y retomando una de las principales obligaciones arriba mencionadas, los contribuyentes tendrán la obligación de autodeterminar los impuestos federales que les

correspondan, ubicándose en el supuesto que señale la legislación fiscal vigente, aplicando el procedimiento específico para llegar a determinar el impuesto correspondiente. Derivado de dicha autodeterminación se puede obtener como resultado un impuesto a cargo o un saldo a favor.

En los casos en que se determine un impuesto a cargo, realizarán el pago en ventanilla bancaria o a través de Internet. Cuando se incumplen estos pagos y la autoridad fiscal los detecta, le genera al contribuyente un crédito fiscal y le exige mediante la cobranza que cumpla con sus obligaciones.

Por el contrario, en los casos en que resulte un saldo a favor se puede optar por solicitar la devolución o la compensación de dicho saldo solicitándolo a la autoridad fiscal.

II.1. Las devoluciones.

¿Qué son las devoluciones?

Las devoluciones son un derecho que tienen los contribuyentes de recuperar las cantidades que hayan pagado indebidamente al SAT, así como la de los saldos a favor que resulten en sus declaraciones, siempre que se hayan determinado correctamente y conforme a lo previsto en las disposiciones fiscales vigentes.

De conformidad con lo establecido en el artículo 22, 22-A, 22-B y 22-C del Código Fiscal de la Federación, las autoridades fiscales están obligadas a devolver a los contribuyentes las cantidades pagadas indebidamente y las que procedan conforme a las leyes fiscales.

En las disposiciones fiscales se señala qué impuestos son sujetos de compensar y en qué casos se puede optar por la devolución del saldo a favor.

Quienes pueden optar por este beneficio son los contribuyentes personas físicas y morales inscritos en el padrón fiscal.

Un caso en el que se puede generar un saldo a favor, es el siguiente:

- Cuando las retenciones que llegan a efectuar son mayores al impuesto que resulta a pagar.

Si el contribuyente obtuvo un saldo a favor, de acuerdo con las disposiciones fiscales se tiene la opción de solicitar la devolución del mismo, deberá cumplir con el procedimiento que señala el Artículo 22 del Código Fiscal de la Federación.

Con fundamento en el Reglamento Interior del Servicio de Administración Tributaria, la resolución de solicitudes de devolución le corresponde a las Administraciones Generales de Auditoría Fiscal y Grandes Contribuyentes, dependiendo de la clasificación del contribuyente que solicita la devolución. Éstas son las encargadas de llevar a cabo el Procedimiento de dictaminación de la resolución de las devoluciones y compensaciones.

La devolución forma parte de la justicia fiscal, ya que si el contribuyente en esta relación jurídico-tributaria tiene una deuda fiscal, pero simultáneamente tiene un saldo a favor, basta con finiquitar esas deudas recíprocas mediante la compensación o devolución

La Autoridad Fiscal se encuentra obligada a devolver las cantidades pagadas indebidamente y las que procedan de conformidad con lo dispuesto en las Leyes Fiscales. Por lo tanto, existe el derecho de los contribuyentes a obtener la devolución de las cantidades que legalmente procedan, en contrapartida, existe la obligación por parte de la Autoridad de pagarlas en tiempo y forma.

Desde su creación en 1997, y ante la necesidad de hacer frente a esta obligación, el SAT ha implementado diversos mecanismos y procesos, así como numerosas herramientas informáticas, para eficientar el proceso de devoluciones y compensaciones.

Las dos áreas operativas responsables del proceso de devolución son:

1. La Administración Central de Devoluciones y Compensaciones de la Administración General de Auditoría Fiscal Federal.
2. La Administraciones Centrales de Fiscalización de Grandes Contribuyentes de la Administración General de Grandes Contribuyentes.

A través de 66 administraciones locales distribuidas a nivel nacional se atienden las solicitudes que los contribuyentes les requieren.

Ante el creciente interés de los contribuyentes en los procesos y en la observación de los plazos establecidos, se hace indispensable proveer a las áreas operativas con herramientas innovadoras para resolver con mayor eficacia y eficiencia las solicitudes de devolución, en tiempo y forma.

Por otra parte, la diversidad de situaciones y problemáticas que se le presentan a las áreas operativas, conjuntamente con la importancia de resolver correctamente las solicitudes de los contribuyentes, hacen necesario el establecer mecanismos y sistemas más asertivos que les permitan resolver en menor tiempo y con mayor calidad.

El objetivo de las áreas operativas encargadas del proceso de devoluciones y compensaciones es proveer a las administraciones locales de políticas y procedimientos conforme a los cuales ejerzan sus atribuciones para controlar los saldos de los contribuyentes, emitir resoluciones dentro del marco legal y normativo vigente, así como homogeneizar su operación a nivel nacional, proporcionando al contribuyente un servicio eficaz, eficiente y de calidad.

Dentro del marco de las leyes dicho objetivo debe cumplir con los plazos que establece el artículo 22 del CFF, por lo tanto se debe devolver o compensar en tiempo y forma.

- Tiempo.

El objetivo por lo que respecta a esta sección es no excederse de los plazos para así evitar la penalización, que se traduce en pago de intereses, demandas de daños y perjuicios, responsabilidades administrativas y mala imagen institucional, entre otros.

Los plazos se definen en función de la contribución de que se trate. Estos plazos oscilan entre 5 y 40 días hábiles, se computan desde el momento en que el contribuyente presenta la solicitud de devolución y hasta que el importe autorizado (en su caso) se encuentra a su disposición.

Para los contribuyentes que dictaminen sus estados financieros por contador público registrado, el plazo para que se efectúe la devolución de saldos a favor de impuestos es de veinticinco días hábiles.

- Forma.

El Código Fiscal de la Federación establece que toda resolución emitida por la Autoridad debe de estar fundada y motivada.

- La fundamentación.

Es el sustento legal en el que se basa la resolución emitida; puede ser el Código Fiscal de la Federación, las Leyes Fiscales (IVA, ISR, IEPS, IMPAC, etc.), los Reglamentos de las Leyes, el Reglamento Interior del SAT, las misceláneas fiscales los decretos o disposiciones de vigencia temporal.

- La motivación.

Se refiere a explicar la causa del por qué se determinó procedente o no procedente una solicitud de devolución; se encuentra estrechamente ligada con la fundamentación, consiste básicamente en interpretar y explicar las disposiciones fiscales aplicables al caso.

El hecho de fundar y motivar correctamente las resoluciones emitidas en relación a las solicitudes de devolución es de vital importancia, principalmente cuando se trate de rechazos; los contribuyentes disponen de medios de defensa que anularán una resolución que no se encuentre debidamente fundada y motivada.

Actualmente la resolución de devoluciones se realiza mediante dos procesos:

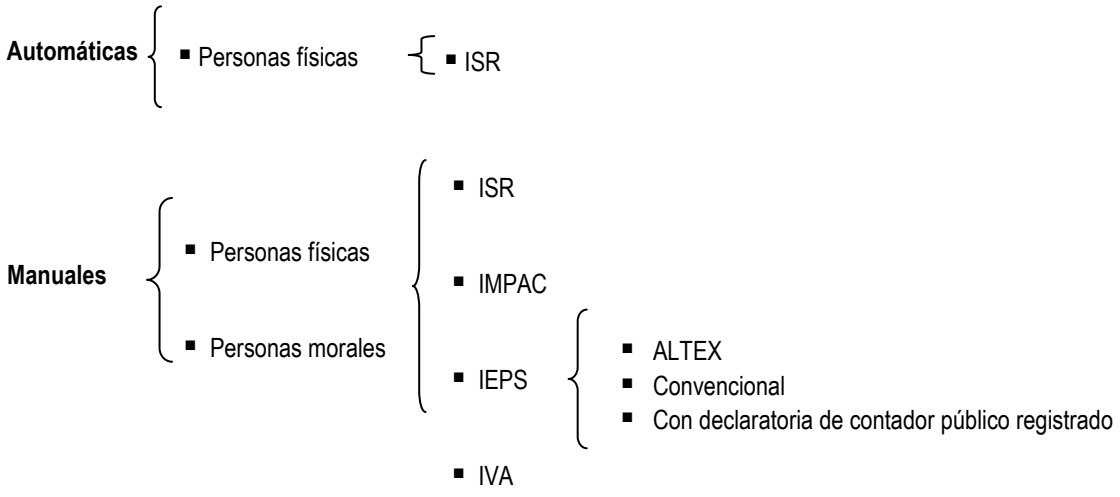


Figura II.1. Procesos para resolución de devoluciones.

Devolución automática: Procede cuando la declaración anual con saldo a favor se presenta dentro del tiempo establecido. No se tiene obligación de realizar ningún trámite adicional para obtener la devolución. Esta opción aplica sólo para personas físicas y para el ISR.

Devolución manual: Ocurre cuando el contribuyente solicita la devolución de algún saldo a favor o pago de lo indebido. Aplica tanto para personas físicas como para personas morales y puede ser por cualquier impuesto, contribución o pago susceptible de devolver. El presente trabajo está enfocado en este tipo de devolución

Las 66 administraciones locales son las responsables de llevar a cabo la dictaminación y resolución de devoluciones y compensaciones.

Los pasos que se siguen para el trámite de una devolución manual son los siguientes:

- Recepción de Trámites.

La recepción de trámites de Devoluciones y Compensaciones está a cargo de la Administración Local de Asistencia al Contribuyente, quien realiza la verificación a los datos y documentación aportada por el contribuyente, registrando la recepción cuando estos sean correctos de acuerdo a los requisitos establecidos para cada caso, entregando el acuse al contribuyente como constancia de la recepción; de no ser así, devuelve la documentación e informa al contribuyente de las inconsistencias detectadas para que las corrija.

- Conformación del Expediente.

Dentro de este módulo, se establecen los procedimientos de recepción, análisis y registro de la documentación anexa a cada promoción de devoluciones y compensaciones; así mismo, se establecen las actividades para la validación de la situación fiscal del contribuyente en general por cada tipo de promoción considerando la información existente en las bases de datos con que cuenta el SAT. Del mismo modo, se indican las actividades para el registro y control del saldo, por último se asigna al dictaminador responsable de resolver cada promoción.

- Requerimientos.

En este apartado, están concentrados los procedimientos relativos a la emisión de requerimientos de devoluciones y compensaciones de documentación faltante y/o adicional, así como cartas invitación y requerimientos derivados de la vigilancia de avisos de compensación y requerimientos de depósitos no efectuados por cuentas bancarias inconsistentes reportadas por TESOFE. Los diversos tipos de requerimientos son direccionados para su notificación a las Unidades de Diligenciación.

- Dictamen.

En este apartado, se establecen de manera general los procedimientos para la dictaminación de devoluciones y compensaciones, de los cuales se derivan actividades específicas para cada tipo de promoción y contribución. Dichas actividades consisten principalmente en llevar a cabo el análisis exhaustivo de la documentación presentada por el contribuyente, así como la aportada en el módulo anterior, la validación aritmética y legal del saldo conforme a su origen, aplicando la legislación y normatividad vigente para cada tipo de contribución; determinando con ello la procedencia o improcedencia de la devolución o compensación.

- Resolución.

Previamente a la resolución que se registra para cada uno de los trámites, en este apartado se tiene considerado realizar procedimientos donde se verifica que los expedientes se encuentren debidamente integrados, que se hayan trabajado bajo el principio de primeras entradas primeras salidas, dentro de los plazos administrativos establecidos, así como que la autorización y registro de resoluciones en el sistema se lleve a cabo.

- Pago.

Dentro del esquema de pago de devoluciones autorizadas intervienen los siguientes actores: la Administración Local de Recaudación, la Administración Central de Devoluciones y Compensaciones, la Tesorería de la Federación, el Banco de México y las Instituciones Bancarias. El envío y recepción de la información entre dichas dependencias se realiza en forma automatizada.

Las actividades que se regulan en este módulo consideran el control de las devoluciones autorizadas a través del envío de emisiones que las mismas realicen a la Administración Central de Devoluciones y Compensaciones, quien gestiona el pago ante la Tesorería de la Federación.

Con base en la información del pago que esta dependencia retroalimenta, es actualizada la información en el ámbito Local y con ello generados los comunicados de pago. Éstos son notificados en los domicilios de los contribuyentes, por los medios de envío que se señalen, a fin de que se cobren directamente en la Institución Bancaria con la presentación del comunicado e identificación del beneficiario.

Cuando el contribuyente señale la cuenta bancaria CLABE el monto autorizado se deposita en ésta, sin generar el comunicado respectivo.

¿Qué sucede cuando se detecta una devolución improcedente?

Una devolución improcedente puede tener diferentes consecuencias, dependiendo de los motivos por los cuales es considerada improcedente; las consecuencias pueden ser ir desde la notificación de un rechazo, hasta un posible juicio penal por defraudación fiscal.

Una vez que se detecta una devolución improcedente, las acciones a seguir dependerán de las circunstancias y motivos que hacen improcedente la devolución; a continuación se presentarán algunos ejemplos.

A. Devolución improcedente por errores aritméticos.

En este supuesto, simplemente se realizará el cálculo de la devolución que, en su caso, realmente le corresponde al contribuyente, y se le pagará el saldo a favor resultante o se le cobrará el saldo a cargo que resulte, en su caso.

En tal supuesto, se le notificará al contribuyente un requerimiento solicitando que aporte la documentación que ampare la determinación del saldo a favor (por ejemplo las constancias de percepciones para comprobar la proporción de subsidio acreditable, o los recibos de honorarios que amparen las deducciones autorizadas, etc.) y, en caso de que no la aporte, se realizarán los cálculos considerando únicamente la documentación aportada.

Consecuencia: determinación de la autoridad o desistimiento de la solicitud.

B. Devolución improcedente por incorrecta aplicación de las Leyes fiscales.

Cuando un contribuyente considera que un artículo de la Ley no le aplica porque fue declarado inconstitucional (ISCAS, ART. 109 LISR, etc.), en ocasiones autodetermina un saldo a favor considerando que no debió de haber pagado ese impuesto solicitando la devolución.

En tal supuesto considerando que una declaratoria de inconstitucionalidad es insuficiente para exentar a un contribuyente del pago de una contribución, se emite un requerimiento solicitando la sentencia en la cual se le haya otorgado el amparo al contribuyente, en caso de que el contribuyente no la presente, se deberá de negar la devolución; en este punto es importante señalar que la resolución deberá de estar debidamente fundada y motivada, ya que en todos los casos el contribuyente demanda la nulidad de este tipo de resoluciones.

Consecuencia: rechazo de la solicitud.

C. Devolución improcedente por detectar documentación apócrifa o falsedad en declaraciones.

Considerando que la Autoridad no cuenta con plena certeza en cuanto a la veracidad o precisión de las manifestaciones hechas por los contribuyentes al momento de resolver una solicitud de devolución, es bastante complicado y poco común detectar una devolución fraudulenta. Lo anterior no significa que no se lleguen a presentar, simplemente que en la actualidad la autoridad que resuelve no cuenta con todos los elementos que permitirían detectar un fraude (acceso a información bancaria, acceso a registros contables, etc.).

Por otra parte, el SAT no cuenta con las facultades necesarias para determinar si se configura o no el supuesto de defraudación fiscal en relación a solicitudes de devolución, únicamente se limita a brindar a la Procuraduría Fiscal los elementos que detectó para que esta actúe en consecuencia.

Existe el caso de terceros que cuentan con un poder notarial que les autoriza a solicitar la devolución en nombre de los beneficiarios de la devolución; en estos casos, por ejemplo, es imposible determinar de quien es la cuenta bancaria que se asentó en la solicitud, por lo tanto es factible que se le pague al tercero (representante legal) sin que el beneficiario de la devolución se entere.

Dependiendo de los alcances del poder notarial, quien debería de presentar la demanda civil es el contribuyente, ya que la Autoridad no tiene las facultades para negarse a pagar una devolución promovida por un representante legal debidamente acreditado.

Consecuencia: rechazo y se turna el expediente a la procuraduría fiscal, para que determinen las acciones a tomar.

La obligación de la autoridad fiscal de devolver saldos a favor prescribe en un plazo de cinco años.

El proceso general de devoluciones se muestra a continuación:

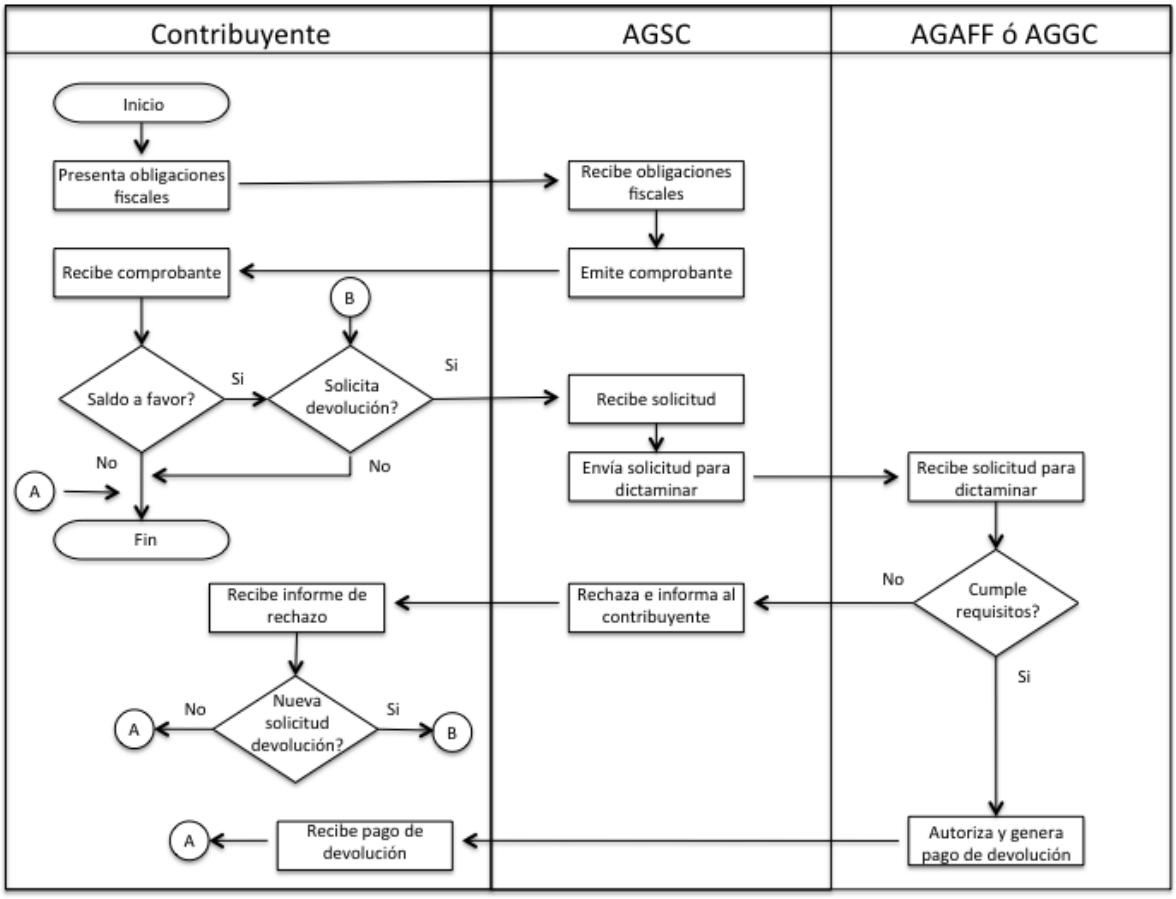


Figura II.2. El proceso general de devoluciones.

El proceso de dictaminar una solicitud de devolución es la misma para todos los contribuyentes, esto puede ocasionar una tardanza excesiva en los trámites ya que se le dá el mismo trato a una devolución de unos cientos de pesos a otras de millones.

El plazo de ley para la resolución debe de cumplirse, de lo contrario el SAT está obligado a pagar intereses por su demora, es necesario tener una forma de discriminar las solicitudes que merecen más análisis a fondo por el riesgo (monto) que significan para el erario público.

Modelo de riesgo.

El riesgo del proceso de devoluciones es: devolver una cantidad cuando no se debió hacerlo, ya que esto ocasionaría una erogación errónea para el fisco, esto puede ocasionarse por muchos factores entre los que podrían estar desde errores administrativos hasta posibles actos de corrupción.

El proceso tiene la discrecionalidad de una dictaminación manual por parte de funcionarios públicos. Esto quiere decir que el dictaminador responsable del asunto puede en última instancia decidir si la devolución procede.

Un modelo predictivo se entrenaría con las solicitudes de devolución históricas, con éstas generará reglas para determinar si se debe pagar una devolución, esto se podría realizar de manera muy rápida lo que ayudaría mucho a resolver a tiempo las solicitudes así como eliminaría la discrecionalidad en las resoluciones.

II.2. La cobranza.

¿Qué es la cobranza?

La cobranza es un proceso que se origina de actos de fiscalización por parte del SAT, éstos se realizan para constatar el cumplimiento de las obligaciones y deberes fiscales del contribuyente.

El artículo 42 del CFF dice:

“Las autoridades fiscales a fin de comprobar que los contribuyentes, los responsables solidarios o los terceros con ellos relacionados han cumplido con las disposiciones fiscales y, en su caso, determinar las contribuciones omitidas o los créditos fiscales, así como para comprobar la comisión de delitos fiscales y para proporcionar información a otras autoridades fiscales, estarán facultadas para:

…

II. Requerir a los contribuyentes, responsables solidarios o terceros con ellos relacionados, para que exhiban en su domicilio, establecimientos o en las oficinas de las propias autoridades, a efecto de llevar a cabo su revisión, la contabilidad, así como que proporcionen los datos, otros documentos o informes que se les requieran.”

Un acto de fiscalización se presenta cuando la autoridad revisa a un contribuyente en el correcto cumplimiento de sus obligaciones, uno de los objetivos es buscar que se tenga una percepción de riesgo y por lo tanto se cumplan con las obligaciones de forma voluntaria y conforme a derecho.

Los contribuyentes dependiendo de su régimen fiscal deben cumplir con declarar, manifestar o cumplir avisos, la autoridad debe revisar desde que se presenten las obligaciones en tiempo como la forma en la cual se declaran los conceptos sean correctos.

Una vez que se presentan las obligaciones en tiempo, se debe de revisar cada uno de los conceptos con declaraciones incorrectas o datos falsos, esto se le conoce como evasión fiscal.

El objetivo que persiguen los actos de fiscalización es reducir al máximo la evasión fiscal, proceder a determinar el monto de las obligaciones omitidas para el cobro y recaudación de los impuestos.

Cuando la autoridad fiscal determina que el contribuyente incumplió sus obligaciones, le genera lo que se llama un crédito fiscal, éste en el código fiscal se define de la siguiente manera.

“Artículo 4. Son créditos fiscales los que tengan derecho a percibir el Estado o sus organismos descentralizados que provengan de contribuciones, de aprovechamientos o de sus accesorios, incluyendo los que deriven de responsabilidades que el estado tenga derecho a exigir de sus servidores públicos o de los particulares, así como aquellos a los que las leyes les den ese carácter y el Estado tenga derecho a percibir por cuenta ajena.”

Es importante mencionar que cuando la obligación fiscal se traduce en una carga líquida y en dinero, será un crédito fiscal, esto quiere decir que éste solo se generará cuando implique un importe a cobrar.

La autoridad encargada del cobro de los créditos fiscales dentro del SAT es la Administración General de Recaudación, es quien administrará los procesos para ejercer las acciones de cobro. La forma en la cual un crédito fiscal se finiquita es cuando el contribuyente cumple con el pago o cuando la ley autoriza extinguir la obligación.

El nacimiento de un crédito fiscal se genera cuando la autoridad ejerce sus facultades de comprobación de las obligaciones del contribuyente y encuentra irregularidades que se traducen en importes a favor del fisco que debe cobrar. La exigibilidad del adeudo tiene una vigencia de 5 años a partir de la determinación que da origen al crédito.

La comprobación de las obligaciones es fundamental para la creación de los créditos fiscales. La autoridad tiene las siguientes facultades para comprobar:

- Visita domiciliaria.
- Revisión de gabinete.
- Revisión de dictámenes, ya sea de estados financieros o de cumplimiento de obligaciones fiscales.
- Inspección de instalaciones, mediciones, bienes y mercancías, así como verificar no sólo el estado y condiciones de los dispositivos permanentes de medición continua de las descargas a la red de drenaje, sino también el volumen de agua tratada.
- Valuación de bienes inmuebles.
- Presenciar la celebración de loterías, rifas, sorteos, concursos, juegos con apuestas y apuestas.
- permitidas de toda especie y verificar los ingresos que perciban.

Una vez determinado un crédito fiscal se siguen procesos de cobro. Al final el crédito fiscal puede terminar por las siguientes situaciones:

PAGO.

Artículos 6°, 20, 21, 31, 65 y 66 del Código Fiscal de la Federación.

El cual debe ser realizado por el contribuyente, representante legal o por terceros (subrogación) en cuanto a las garantías, preferencias y privilegios sustanciales. El pago debe efectuarse en el lugar, la fecha y la forma que indique la ley o en su defecto la reglamentación.

Los pagos anticipados deben ser expresamente dispuestos por la ley. Los contribuyentes pueden optar por prórrogas (pago diferido) y facilidades de pago (pago en parcialidades) que deben solicitarse antes del vencimiento del plazo para el pago y sólo podrán ser concedidas cuando a juicio de la autoridad se justifiquen las causas que impiden el cumplimiento normal de la obligación. La decisión de denegar las prórrogas y facilidades por parte de la autoridad no admitirá recurso alguno.

COMPENSACIÓN.

Se compensarán de oficio o a petición de parte los créditos líquidos y exigibles del contribuyente por concepto de tributos, con las deudas tributarias liquidadas por aquel y no observadas, comenzando por los más antiguos y aunque provengan de distintos tributos, siempre que sean administrados por la misma autoridad administrativa.

Cuando la administración determine nuevas obligaciones, el contribuyente podrá compensarlas por créditos que tenga por los mismos periodos fiscales invocados con anterioridad, aunque estuviesen preescritos. Igual derecho tendrá la administración en las situaciones análogas producidas a raíz de reclamaciones del contribuyente.

Los créditos que sean líquidos y exigibles del contribuyente por concepto de tributos podrán ser cedidos a otros contribuyentes y responsables al solo efecto de ser compensados con deudas tributarias que tuviese el cesionario en el mismo órgano administrativo.

CONDONACION.

La condonación del pago de los tributos solo puede ser condonada o remitida por leyes dictadas con alcance general. Además las obligaciones, así como los intereses y las multas, sólo pueden ser condonadas por resolución administrativa en la forma y condiciones que la ley establezca.

PRESCRIPCIÓN.

Conforme al artículo 146 de Código Fiscal de la Federación los créditos fiscales prescriben en 5 años. El término de la prescripción se inicia a partir de la fecha en que el pago pudo ser legalmente exigido y se podrá poner como excepción en los recursos administrativos.

La prescripción se interrumpe con cada gestión de cobro que el acreedor notifique o haga saber al deudor o por el reconocimiento expreso o tácito de éste respecto de la existencia del crédito.

La suspensión procede cuando no sea legalmente posible exigir el pago; cuando se suspenda el procedimiento administrativo y su efecto será que no se iniciará o correrá el plazo, concluida la suspensión, se inicia el plazo o se continúa el ya indicado.

CANCELACION.

Consiste la cancelación de un crédito por insolvencia del deudor o incosteabilidad en el cobro. La Ley de Ingresos de la Federación señala:

Artículo 15. Se faculta a las autoridades fiscales para que lleven a cabo la cancelación de los créditos fiscales cuyo cobro les corresponda efectuar, en los casos en que exista imposibilidad práctica de cobro. Se considera que existe imposibilidad práctica de cobro, entre otras, cuando los deudores no tengan bienes embargables, el deudor hubiera fallecido o desaparecido sin dejar bienes a su nombre o cuando por sentencia firme hubiera sido declarado en quiebra por falta de activo.

El proceso general para la generación de un crédito fiscal se muestra a continuación:

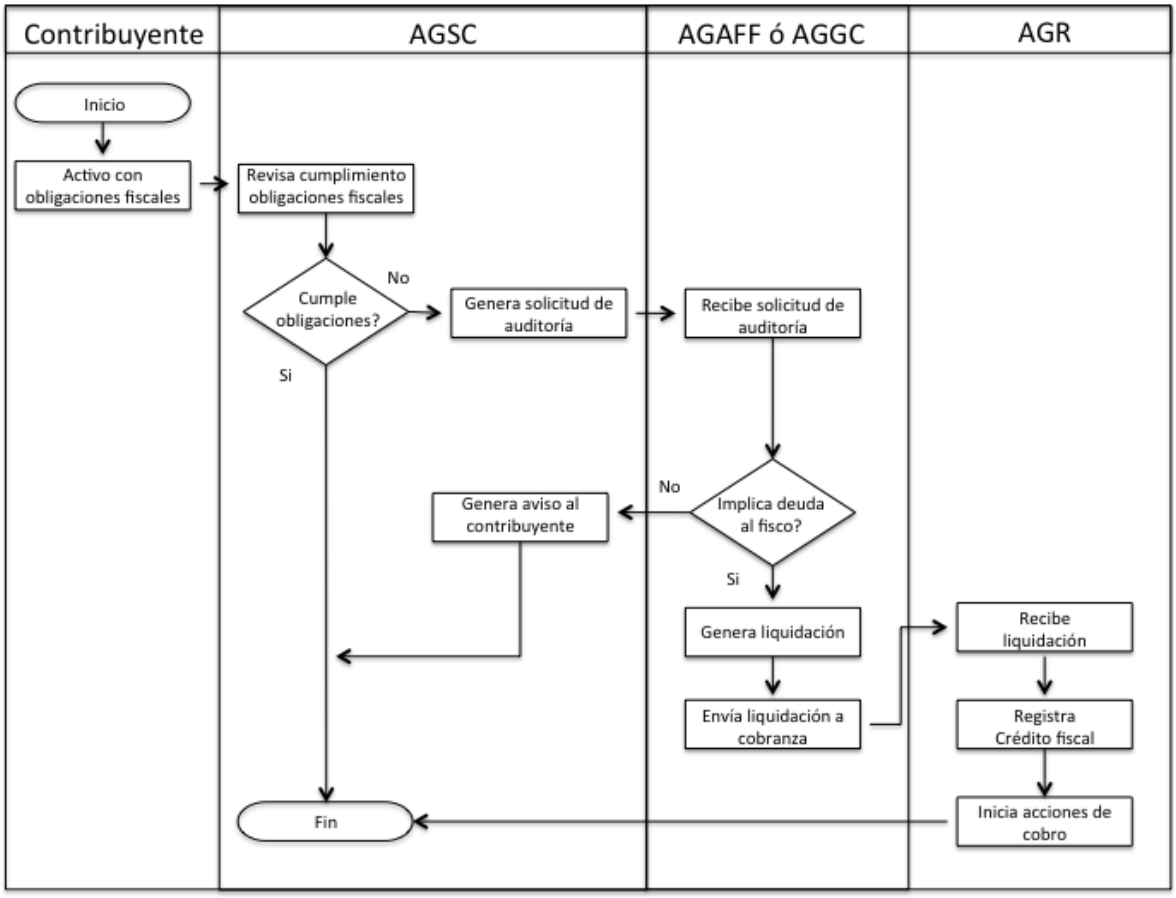


Figura II.3. El proceso general para generar un crédito fiscal.

Al generarse un crédito fiscal la autoridad fiscal está obligada a realizar todos los procesos que la ley le permite para poder cobrarlo, existen procesos generales que se realizan para llevar a cabo estas acciones.

1. Notificación.

Es el medio establecido a través del cual la autoridad fiscal da a conocer a los contribuyentes, a los responsables solidarios o a terceros, el contenido de un acto administrativo derivado de la fiscalización a efecto de que estén en posibilidad de cumplirlo o de impugnarlo.

El acto administrativo debe ser notificado por escrito y sustentado con los correspondientes fundamentos legales en los que la autoridad competente se soporte, el documento con estas características se denomina acta de notificación. Las disposiciones legales que regulan las notificaciones en materia fiscal son del 134 al 140 del CFF.

Los tipos de notificación son:

Personales o por correo certificado o electrónico con acuse de recibo. Se lleva a cabo a través de una visita domiciliaria encontrando al contribuyente o al representante legal, también puede ser por correo certificado, el caso de correo electrónico, será el documento digital con Firma Electrónica Avanzada que transmita el destinatario al abrir el documento digital que le hubiera sido enviado. La fecha de notificación se considerará el día que ésta se reciba.

Estrados. El documento que se deba notificar se fija durante 15 días consecutivos en las oficinas de la autoridad que notifica, en un sitio abierto al público, y dando a conocer el documento durante el mismo plazo, en la página electrónica de la autoridad fiscal. Este tipo de notificación se lleva a cabo cuando el contribuyente desaparezca, se oponga a la notificación, desocupe el local donde tenga su domicilio fiscal sin presentar aviso de cambio de domicilio. Se considerará la notificación después de 15 días después de fijado o publicado el aviso.

Edictos. Se publicará un resumen de los actos que se notifican durante tres días consecutivos en el Diario Oficial de la Federación; por un día en uno de los periódicos de mayor circulación, y durante quince días consecutivos en la página electrónica de la autoridad fiscal. Esto se realiza cuando el contribuyente haya fallecido y se desconozca al representante de la sucesión, hubiese desaparecido o el domicilio no se encuentre en territorio nacional. La fecha de notificación será la de la última publicación en el Diario Oficial de la Federación o en la página electrónica, según sea el caso

Correo ordinario o telegrama. También se puede llevar a cabo este tipo de notificación cuando los actos no se hayan llevado a cabo por los anteriores. También se considerará como notificado la fecha en la cual se recibe el documento.

Después de cumplida la fecha de notificación el contribuyente cuenta con 45 días para el pago del crédito fiscal o para presentar un medio de defensa cuando considere que la resolución no se llevó a cabo conforme a la ley.

2. Medios de defensa.

Los medios de defensa que puede promover el contribuyente sobre resoluciones definitivas son:

Recursos de revocación. Es un acto de naturaleza administrativa ya que no implica una función plenamente jurisdiccional por parte de una autoridad independiente e imparcial a la determinación del acto. Es la misma autoridad (SAT) quien revisa la legalidad de lo que ella misma emitió sobre el

contribuyente. Esto corresponderá a las Administraciones Locales Jurídicas (dependiente de la Administración General Jurídica) que corresponda al domicilio fiscal del contribuyente.

Este recurso es regulado por los artículos 116 al 133-A del Código Fiscal de la Federación. Se podrá promover dentro de los 45 días hábiles después de la notificación. La autoridad deberá resolver y notificar la resolución en un plazo no mayor a 3 meses contados a partir de la interposición del recurso.

Juicios de nulidad. El contribuyente puede optar también por acudir directamente al Tribunal Federal de Justicia Fiscal y Administrativa para demandar la nulidad del acto o resolución, también podrá interponer un juicio de nulidad ante el citado Tribunal. Podrá presentarse dentro de los 45 días después de la notificación.

Juicio de amparo. Pueden ser directos o indirectos dependiendo del tipo de resolución que se le generó al contribuyente, el primero es contra sentencias definitivas y el segundo es contra no definitivas.

La resolución de cualquier medio de defensa puede derivar que el créditos fiscal sea cancelado cuando la autoridad determine que no debió generarse o mantenerse firme cuando el fallo sea en contra del contribuyente por lo que se seguirán las acciones de cobro.

Es importante mencionar que cuando el contribuyente interpone un medio de defensa o solicita pagar con alguna facilidad debe de garantizar el interés fiscal.

3. Garantías.

La garantía del interés fiscal es una figura jurídica dentro de la administración tributaria cuyo fin es asegurar el pago por parte del contribuyente respecto de un crédito fiscal que se encuentra en una situación (por ejemplo medios de defensa) en la que se espera una resolución.

También cuando el contribuyente opta por tener facilidades para el pago la autoridad exige una garantía para poder otorgarle ese beneficio.

Esto se encuentra dentro del CFF en sus artículos 141 y 142.

Las 6 formas de garantizar el interés fiscal son:

1. Depósito en dinero, carta de crédito o Garantía Financiera.
2. Prenda o hipoteca.
3. Fianza.
4. Obligación solidaria de un tercero.
5. Embargo en vía administrativa.
6. Títulos valor o cartera de créditos del propio contribuyente.

El objetivo de la garantía es que la autoridad fiscal asegure el cobro del crédito fiscal aún y cuando éste al final pueda ser desestimado por alguna autoridad competente.

4. Cobro persuasivo.

Es un proceso dentro de la cobranza que intenta motivar en el contribuyente a través de estructuras amigables el pago de sus créditos fiscales evitando con ello el llegar a otro tipo de mecanismos de los cuales dispone la autoridad para cobrar como lo es el Procedimiento Administrativo de Ejecución que por su naturaleza resulta mucho más agresivo (coactivo) para el contribuyente.

El cobro persuasivo también implica un menor gasto para la autoridad debido a que invierte mucho menos dinero en el proceso. Los canales de comunicación con el contribuyente para este proceso son los siguientes:

- A través del correo postal. Cartas de invitación enviadas al domicilio del contribuyente o a domicilios alternos con su respectiva línea de captura para realizar el pago.

- De forma electrónica por correo electrónico con formato para pago de contribuciones federales que contiene la línea de captura.

- Mediante llamadas telefónicas en el Centro Telefónico de Cobranza, donde se invita a regularizar los adeudos, también se le envian al contribuyente los formatos necesarios para el pago de sus adeudos.

5. Cobro coactivo.

Como ya se ha mencionado en la Constitución Política de los Estados Unidos Mexicanos se especifica que es obligación de los mexicanos pagar impuestos, sin embargo cuando estos no los pagan de manera voluntaria el Estado se los cobra de manera forzosa, mediante la facultad económico coactiva que tiene concedida.

Este proceso dentro de la cobranza se conoce como el Procedimiento Administrativo de Ejecución (PAE). Se trata del medio coercitivo del cual disponen las autoridades para exigir el pago de los créditos fiscales que no hubieren sido cubiertos o garantizados dentro de los plazos señalados por la ley fiscal. El fundamento constitucional es el artículo 22 de la Constitución Política de los Estados Unidos Mexicanos.

El PAE se lleva a cabo para hacer efectivo un crédito fiscal exigible y el importe de sus accesorios legales, las autoridades fiscales están facultadas para requerir de pago al deudor y, en caso de que este no pruebe en el acto haberlo efectuado, deberán proceder a embargar bienes suficientes para, en su caso, rematarlos, enajenarlos fuera de subasta o adjudicarlos a favor del fisco o embargar negociaciones con todo lo que de hecho y por derecho les corresponda, a fin de obtener, mediante la intervención de ellas, los ingresos necesarios que permitan satisfacer los créditos fiscales y los accesorios legales. (Artículo 151 del CFF).

Los abogados tributarios son los encargados de realizar esta actividad observando lo establecido en el CFF ya que todos los actos que realicen deben de cumplir las formalidades de ley para que al final se llegue a buen término el cobro de los créditos fiscales.

El PAE se desarrolla, a través de una serie de actos procedimentales, que tienen el carácter de actos administrativos.

El primer acto administrativo es el *Mandamiento de ejecución*. Es la resolución que dicta la autoridad recaudadora en la que se encuentre radicado el crédito fiscal, en la que se ordenará que se requiera al deudor para que acredite haber efectuado el pago en el mismo acto de la diligencia de requerimiento, con el apercibimiento que de no hacerlo, se le embargarán bienes suficientes para hacer efectivo el crédito fiscal y sus accesorios, para en su caso, obtener mediante el remate de los bienes, el importe de los créditos adeudados.

El siguiente acto administrativo es el *Requerimiento de pago*. Es la diligencia por medio de la cual, las autoridades fiscales exigen el pago del crédito fiscal no cubierto o garantizado en los plazos legales establecidos. Debe de realizarse en el domicilio del deudor, se debe de notificar en forma personal y levantar un acta con los hechos realizados.

Existen diversos tipos de Requerimiento de Pago, entre ellos, los siguientes:

a) Requerimiento de Pago por no haber pagado o garantizado dentro del término legal.

   Está regulado en los artículos 145 primer párrafo, 151 primer párrafo, 65 y 144 del Código Fiscal de la Federación.

b) Requerimiento de Pago por cese de la autorización del pago a plazos.

   Está regulado en los artículos 151 último párrafo y 66-A Fracción IV del Código Fiscal de la Federación.

c) Requerimiento de Pago por errores aritméticos o por falta de presentación de la declaración.

   Está regulado en los artículos 151 último párrafo y 41 Fracción I del Código Fiscal de la Federación.

d) Requerimiento de Pago por concepto de multas judiciales o reparación del daño.

   Se implementa para hacer efectivas las multas judiciales y la reparación del daño que determinan las autoridades del Poder Judicial de la Federación.

El PAE continúa con el siguiente acto administrativo: *El embargo*. El cual, se lleva a cabo con posterioridad al requerimiento de pago y representa en sí, la medida coactiva con que cuenta el Estado para allegarse de los elementos necesarios para hacer cumplir sus determinaciones; en materia tributaria, para obtener la recuperación de los créditos fiscales mediante la enajenación de los bienes que hayan sido afectados mediante esta medida.

Tiene por objeto la recuperación de créditos, mediante el secuestro o aseguramiento de bienes propiedad del contribuyente o deudor, para en su caso, rematarlos, enajenarlos fuera de subasta o adjudicarlos a favor del fisco.

En el CFF, en su artículo 151, dispone que las autoridades fiscales, para hacer efectivo un crédito fiscal exigible y el importe de sus accesorios legales, requerirán de pago al deudor y, en caso de que éste no pruebe en el acto haberlo efectuado, procederán de inmediato como sigue:

1. A embargar bienes suficientes para, en su caso, rematarlos, enajenarlos fuera de subasta o adjudicarlos a favor del fisco.

2. A embargar negociaciones con todo lo que de hecho y por derecho les corresponda, a fin de obtener, mediante la intervención de ellas, los ingresos necesarios que permitan satisfacer el crédito fiscal y los accesorios legales.

La finalidad inmediata del embargo es la de garantizar con bienes suficientes el interés fiscal del Estado para que, llegado el momento, con el producto de su enajenación se alcance la recuperación de los créditos fiscales que tiene a su favor por parte de los particulares.

Los actores que intervienen en el embargo son:

El deudor: Puede ser en su defecto, la persona con quien se entienda la diligencia para el caso de que no hubiere atendido el citatorio en caso de haber precedido.

Los testigos: Los cuales podrán ser nombrados por la persona con quien se entienda la diligencia de embargo.

El ejecutor: El cual es el elemento necesario que lleva a cabo la diligencia.

La autoridad municipal o local: La que se encuentra en la circunscripción de los bienes, si el requerimiento o la notificación se hicieron por edictos.

El ejecutor, al momento de llevar a cabo la diligencia de embargo, deberá cumplir puntualmente con cada una de las formalidades a que se refiere el artículo 137 del Código Fiscal de la Federación.

Bienes susceptibles de embargo. En el artículo 155 del Código Fiscal de la Federación, manifiesta que corresponde al deudor en primer término la designación de los bienes sobre los cuales deba realizarse el embargo, para lo cual, dicha norma establece el orden en el cual preferentemente deban señalarse los bienes, éstos deben ser de fácil realización o venta, y el orden será el siguiente:

- Dinero, metales preciosos y depósitos bancarios.

- Acciones, bonos, cupones vencidos, valores mobiliarios y en general créditos de inmediato y fácil cobro a cargo de entidades o dependencias de la Federación, Estados y Municipios y de instituciones o empresas de reconocida solvencia.

- Bienes muebles no comprendidos en las fracciones anteriores.

- Bienes inmuebles. En este caso el deudor o la persona con quien se entienda la diligencia deberá manifestar, bajo protesta de decir verdad, si dichos bienes reportan cualquier gravamen real, embargo anterior, se encuentran en copropiedad o pertenecen a sociedad conyugal alguna.

Existen tres tipos de embargo:

Embargo definitivo: Tiende a satisfacer una pretensión ejecutiva, una vez que la obligación es plenamente exigible.

Embargo en la vía administrativa: Este tipo de embargo, en términos del artículo 141, Fracción V del Código Fiscal de la Federación, se lleva a cabo a petición del deudor, cuando solicita garantizar el interés fiscal a través de dicha figura con motivo de la solicitud para cubrir sus adeudos en pago diferido o parcialidades o por haber promovido o interpuesto algún medio de defensa.

Embargo precautorio: Este tipo de embargo, conforme a lo dispuesto por el Artículo 145 del Código Fiscal de la Federación se podrá practicar sobre los bienes o la negociación del contribuyente, cuando el crédito fiscal no sea exigible. Se lleva a cabo en los siguientes casos:

I. El contribuyente se oponga u obstaculice la iniciación o desarrollo de las facultades de comprobación de las autoridades fiscales o no se pueda notificar su inicio por haber desaparecido o por ignorarse su domicilio.

II. Después de iniciadas las facultades de comprobación, el contribuyente desaparezca o exista el riesgo inminente de que oculte, enajene o dilapide sus bienes.

III. El contribuyente se niegue a proporcionar la contabilidad que acredite el cumplimiento de las disposiciones fiscales, a que está obligado.

IV. El crédito fiscal no sea exigible pero haya sido determinado por el contribuyente o por la autoridad en el ejercicio de sus facultades de comprobación, cuando a juicio de ésta exista el peligro inminente de que el obligado realice cualquier maniobra tendiente a evadir su cumplimiento.

V. Se realicen visitas a contribuyentes, con locales, puestos fijos o semifijos en la vía pública y dichos contribuyentes no puedan demostrar que se encuentran inscritos en el Registro Federal de Contribuyentes, ni exhibir los comprobantes que amparen la legal posesión o propiedad de las mercancías que vendan en esos lugares.

El embargo se puede realizar a:

Bienes muebles: De conformidad con el Código Civil Federal, los bienes son muebles por su naturaleza o por determinación de la Ley. Son bienes muebles por naturaleza, los cuerpos que pueden trasladarse de un lugar a otro, ya se muevan por sí mismos, ya por efecto de una fuerza exterior.

Bienes inmuebles: Son los mencionados en el artículo 750 del Código Civil Federal:

I. El suelo y las construcciones adheridas a él.

II. Las plantas y árboles, mientras estuvieren unidos a la tierra, y los frutos pendientes de los mismos árboles y plantas mientras no sean separados de ellos por cosechas o cortes regulares.

III. Todo lo que esté unido a un inmueble de una manera fija, de modo que no pueda separarse sin deterioro del mismo inmueble o del objeto a él adherido.

IV. Las estatuas, relieves, pinturas u otros objetos de ornamentación, colocados en edificios o heredados por el dueño del inmueble, en tal forma que revele el propósito de unirlos de un modo permanente al fundo.

V. Los palomares, colmenas, estanques de peces o criaderos análogos, cuando el propietario los conserve con el propósito de mantenerlos unidos a la finca y formando parte de ella de un modo permanente.

VI. Las máquinas, vasos, instrumentos o utensilios destinado por el propietario de la finca directa y exclusivamente, a la industria o explotación de la misma.

VII. Los abonos destinados al cultivo de una heredad, que estén en las tierras donde hayan de utilizarse, y las semillas necesarias para el cultivo de la finca.

VIII. Los aparatos eléctricos y accesorios adheridos al suelo o a los edificios por el dueño de éstos, salvo convenio en contrario.

IX. Los manantiales, estanques, aljibes y corrientes de agua, así como los acueductos y las cañerías de cualquiera especie que sirvan para conducir los líquidos o gases a una finca o para extraerlos de ella.

X. Los animales que formen el pie de cría en los predios rústicos destinados total o parcialmente al ramo de ganadería; así como las bestias de trabajo indispensables en el cultivo de la finca, mientras están destinadas a ese objeto.

XI. Los diques y construcciones que, aun cuando sean flotantes, estén destinados por su objeto y condiciones a permanecer en un punto fijo de un río, lago o costa.

XII. Los derechos reales sobre inmuebles.

XIII. Las líneas telefónicas y telegráficas y las estaciones radiotelegráficas fijas.

Embargo de la negociación: Es el Embargo de todo lo que de hecho y por derecho le corresponda a la empresa o negocio, el ejecutor deberá:

- Levantar inventario de todos aquellos activos que la componen,
- Anexar a la diligencia la relación de activos,
- Acta constitutiva que contenga folio mercantil, y Copia del poder notarial.

El remate.

Es el acto administrativo conformado por el conjunto de actos jurídicos a través del cual, las autoridades fiscales realizan la enajenación forzada en subasta pública, de los bienes embargados a los deudores para con el producto obtenido de la venta, cubrir los créditos fiscales a favor del erario federal.

Por medio de la subasta pública, la Administración Pública actúa solo como vendedora de bienes para hacer efectivo el importe de créditos fiscales a cargo de los particulares propietarios de los bienes subastados.

El producto obtenido del remate, enajenación o adjudicación de los bienes al fisco, deberá aplicarse a cubrir el crédito fiscal en el orden que establece el artículo 20 del Código Fiscal de la Federación, siendo éste el siguiente:

- Gastos de ejecución.
- Recargos.
- Multas.
- La indemnización a que se refiere el séptimo párrafo del artículo 21 del CFF.

El proceso general de cobranza de créditos fiscales es el siguiente:

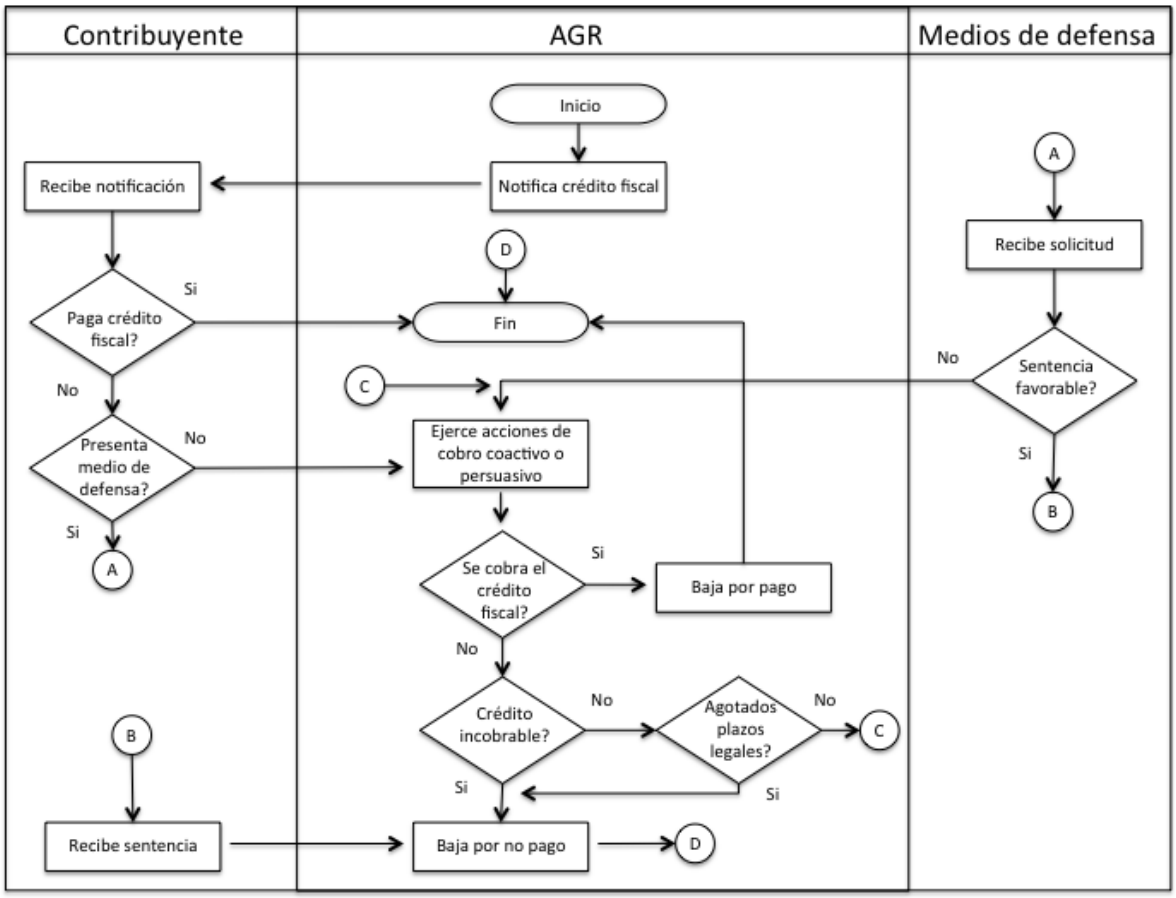


Figura II.4. El proceso general de cobranza.

El contribuyente tiene dos opciones una vez que es notificado, pagar o presentar un medio de defensa, el SAT tiene varios procesos por hacer, como son la notificación, las acciones de cobro, dar de baja los créditos, todos estos procesos necesitan tener ayuda sistematizada para la cobranza.

El tener información adecuada que permita encaminar las acciones de cobro hacia el pago del crédito fiscal sería lo más importante para el área, poder discernir entre las mejores acciones de cobro para determinado contribuyente sería crucial para la cobranza.

Modelo de riesgo.

Un modelo de riesgo predictivo para cobranza podría ayudar en lo siguiente:

- Tener una probabilidad de cobro individual por crédito fiscal.
- Identificar los contribuyentes que son más propensos a no pagar los créditos.
- Generar estrategias de cobro anticipadas de acuerdo a resultados del modelo.
- Generar una aproximación de lo que se puede cobrar de la cartera de créditos fiscales.

II.3. Recapitulación.

Dentro de las instituciones como las empresas es muy importante lo que es sus ingresos y egresos, ambos determinan en mucho el funcionamiento ya que determinan el balance que se tiene en la organización.

Por esta razón se eligieron la cobranza y las devoluciones, dos áreas dentro del SAT que precisamente representan ingresos y egresos.

Las devoluciones son los egresos, la institución por ley tiene que devolver los saldos a favor que el contribuyente obtenga de los diferentes impuestos a los que está obligado a tributar. Es muy importante que esto se realice con eficiencia y celeridad.

Como ocurre en muchas organizaciones se puede aplicar el principio de Pareto (o regla del 80-20), pueden existir muchas solicitudes de devolución (80%) y que en egreso representan poco (20%) importe.

Entonces si se genera para este conjunto de contribuyentes, una herramienta que en base a un modelo predictivo que aprendiendo de la historia indique rápidamente si la solicitud se aprueba o se rechaza, entonces podrá quedar tiempo suficiente para que el 20% restante de las solicitudes que representan el 80% del importe, puedan ser dictaminadas con mucho cuidado para no cometer errores y cumplir en tiempos que marca la ley.

La cobranza son los ingresos, importes que el contribuyente debe al fisco debido a errores en sus obligaciones fiscales o incluso a problemas de evasión fiscal. Sería muy importante para la autoridad poder saber qué contribuyentes pagarán sus adeudos y cuáles no, esto para poder tener una segmentación de los mismos y así aplicarles diferentes esquemas de cobranza.

Un modelo de este tipo puede también encontrar los caminos exitosos y los que no lo son, lo cual ayudaría a la autoridad a identificar hacia donde debe de dirigir la cobranza para que un contribuyente pase de una baja probabilidad de cobro a una alta.

La información fiscal es confidencial, por lo que todo lo que se muestra en el trabajo son solo datos ficticios e información pública que se puede encontrar en portales de internet, por ejemplo www.sat.gob.mx.

Capítulo III. Entendimiento y preparación de datos.

Después de entender el problema que se enfrenta, es fundamental entender y preparar los datos, es importante saber dónde se almacenan, en qué estructuras de datos, etc., para poder manipularlos.

Estos dos procesos dentro de la minería de datos son muy importantes y donde en ocasiones se lleva una mayor cantidad de tiempo dentro de un proyecto. También se le conoce como realizar el ETL (Extract, Transform and Load, extracción, transformación y carga por sus siglas en inglés) de la información.

Para este trabajo se tienen datos de prueba fictícios que mostrarán información de ejemplo con datos no reales ya que la información tributaria es muy delicada y no se podría mostrar públicamente porque se estaría violando el secreto fiscal.

El software de minerìa de datos que se usará para el trabajo es SPSS Clementine 12.0, es uno de los mineros más reconocidos en el mercado y cuenta con todas las herramientas necesarias para el entendimiento, preparación y modelado.

La metodología que se utiliza se basa en CRISP – DM (CRoss Industry Standard Process for Data Mining por sus siglas en inglés) que es la más utilizada por los principales desarrolladores de modelos de minería de datos. Se basa en 6 fases principales a desarrollar:

1. Entendimiento del negocio: Fue lo que se hizo en el capítulo 2, entender el negocio y el problema a resolver.
2. Entendimiento de los datos: Es el paso que sigue en este capítulo.
3. Preparación de los datos: Más adelante en este capítulo.
4. Modelado: La parte de la construcción de modelos que se realizará en el siguente capítulo.
5. Evaluación: Donde se evalúan los resultados del modelado para obtener el mejor modelo. Se revisará en el capítulo 4.
6. Implementación: La puesta en producción del modelo ganador. Este punto no se tratará en el trabajo pero será mencionado una vez que se encuentren los mejores modelos para el trabajo.

Estas etapas son cíclicas (Figura III.1.), lo que quiere decir que en cualquier punto se puede regresar a uno previo para resolver algún problema dentro del proceso.

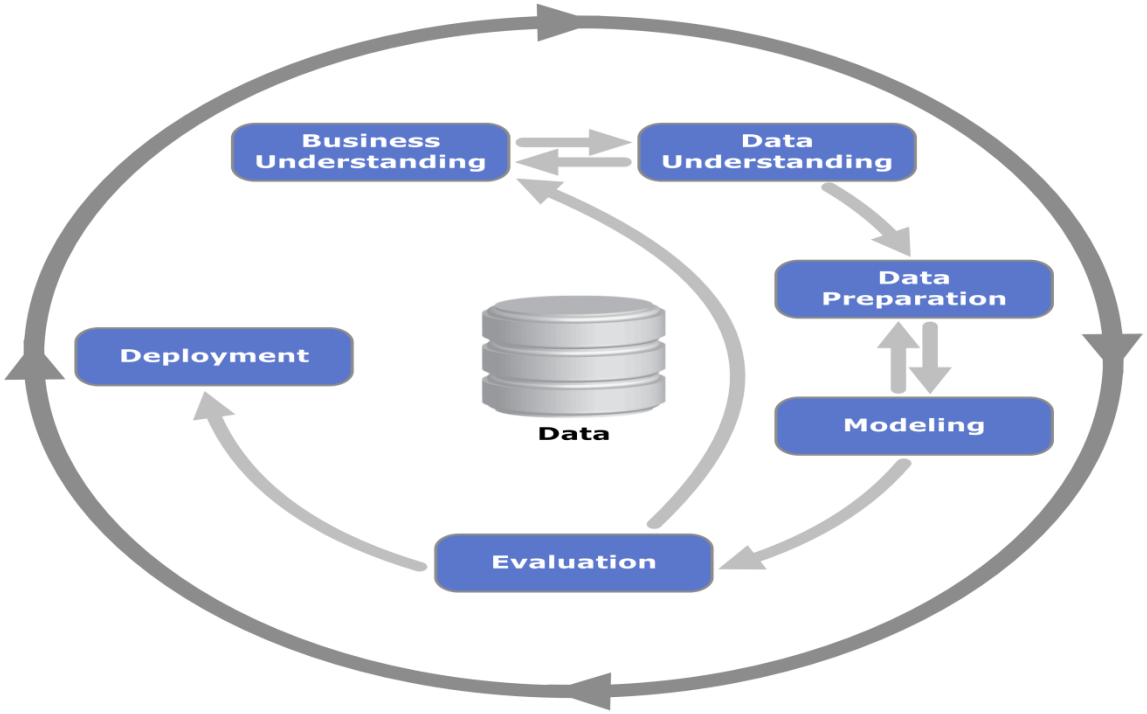


Figura III.1. Flujo de las fases en la metodología CRISP-DM.

El objetivo que se busca en el entendimiento y preparación de los datos es generar una vista minable.

Normalmente la información a trabajar se tiene en tablas bajo un diseño normalizado de bases de datos, esto es, la estructura está diseñada de manera que se optimice el almacenamiento de la información utilizando para esto la construcción de tantas tablas transaccionales y de catálogo necesarias para guardar los datos.

Para preparar la información para un modelo predictivo de minería de datos se debe generar un proceso inverso a la normalizaciòn de bases de datos, al final se tiene que llegar a una sola tabla con todas las variables necesarias para poder generar los modelos. Esta tabla se le conoce como vista minable.

La vista minable se compone de 3 principales tipos de campos o variables:

1. Campos de información. Son aquellos que tienen la información necesaria para poder asignar el resultado al sujeto de estudio, en este caso nuestro sujeto al cual calificará los

modelos es el RFC, éste no servirá como variable de entrada o independiente pero al final será la llave para relacionar los resultados a otros datos dentro de un sistema.

2. Campos de análisis. Son todos aquellos que puedan aportar información para predecir los resultados, pueden ocuparse como están en la BD o transformarse para obtener datos adicionales, por ejemplo, una fecha no debe utilizarse dentro del modelo pero tal vez si el año o el mes o la cantidad de días que han transcurrido.

3. Campo objetivo. Es el campo que se está buscando predecir, puede ser que éste deba transformarse para poder ser utilizado de acuerdo al algoritmo utilizado, por ejemplo si la variable fuera cadena pudiera convertirse en número para poderlo utilizar en una regresión lineal.

La información de las tablas fue convertida a archivos txt para el manejo más práctico con la herramienta de minería de datos.

III.1. Entendimiento de los datos.

Se trata de la primera aproximación que se tiene a la información con la cual se va a trabajar. Una vez que se entendió el problema, hay que revisar cómo está estructurada la información con la que se va a trabajar.

Dentro de una organización se puede almacenar los datos en diferentes tipos de repositorios desde estructurados como bases de datos como no estructurados como archivos planos.

Generalmente en organizaciónes grandes como lo es el SAT, la información se almacena en bases de datos robustas donde la información se guarda para su uso posterior.

Un primer paso para el entendimiento sería revisar los diccionarios de la base de datos que contienen la información correspondiente de cobranza y devoluciones.

III.1.1. Devoluciones.

Para devoluciones la definición de las tablas donde se podría guardar la información sería:

Registro de solicitudes: Registro		
Nombre	Definición	Descripción
Rfc	char(13)	Registro federal de contribuyentes
nomb_contrib	char(255)	Nombre del contribuyente
no_solicitud	int(10)	Número de solicitud de devolución
cve_impuesto	int(2)	Clave de impuesto de la devolución
cve_local	int(2)	Clave de local de recaudación
Importe	real(12,2)	Importe de la solicitud de devolución
fecha_solicitud	Date	Fecha de solicitud

Tabla III.1. Definición tabla de devoluciones.

Se tiene una tabla llamada registro donde se lleva el control de las solicitudes actuales e históricas de las devoluciones.

Registro de resoluciones: Resoluciones		
Nombre	Definición	Descripción
no_solicitud	int(10)	Número de solicitud de devolución
cve_estatus	int(2)	Clave de estado de la solicitud
Importe	real(12,2)	Importe autorizado de la solicitud
fecha_estado	Date	Fecha de estado

Tabla III.2. Definición tabla de resoluciones.

Una segunda tabla donde se registran las resoluciones a las solicitudes de devolución.

Al revisar la definición de ambas tablas se puede saber que ambas están relacionadas por el número de solicitud de devolución (no_solicitud) ya que este campo es común en las dos, se puede interpretar que en la primera se registran todas las solicitudes y en la segunda se tiene el estatus tanto de las solicitudes actuales como las históricas.

Podemos ver que también se tiene un campo importe, éste no necesariamente es el mismo, uno es con el que el contribuyente solicita la devolución y el otro es el que se autoriza a devolver. Incluso uno podría ser diferente del otro cuando al final se autoriza un monto menor al solicitado por algún tipo de inconsistencia en las deducciones.

También se observa que tienen campos con claves que describen los impuestos, los estatus y las locales de recaudación.

Es importante también saber el significado de las claves por lo que a continuación se muestra la definición de los catálogos:

Catálogo de impuestos: Cat_impuesto		
Nombre	Definición	Descripción
cve_impuesto	int(2)	Clave de impuesto de la devolución
desc_impuesto	char(50)	Descripción del impuesto

Tabla III.3. Definición tabla catálogo de impuestos.

Una tabla con los catálogos de los impuestos por los cuales se está solicitando la devolución.

Catálogo de estatus: Cat_estatus		
Nombre	Definición	Descripción
cve_estatus	int(2)	Clave de estado de la solicitud
desc_estatus	char(50)	Descripción de estatus

Tabla III.4. Definición tabla catálogo de estatus.

El catálogo de los estatus por los que pasa la solicitud de devolución.

Catálogo de locales: Cat_local		
Nombre	Definición	Descripción
cve_local	int(2)	Clave de local de recaudación
desc_local	char(50)	Descripción de local

Tabla III.5. Definición tabla catálogo de locales.

El catálogo de las locales de recaudación que tienen bajo su responsabilidad la resolución de las solicitudes.

Una vez que se tienen las tablas con su definición, es importante revisar el diagrama entidad – relación de éstas para confirmar las relaciones entre los datos.

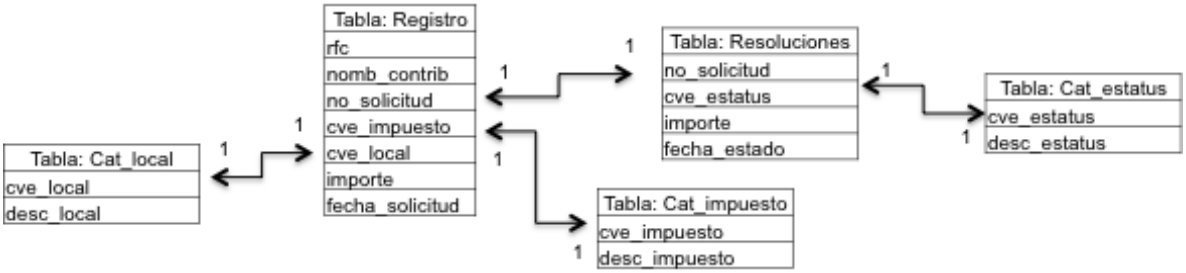


Figura III.2. Diagrama entidad relación de las tablas de devoluciones.

El diagrama entidad relación nos muestra lo que intuitivamente se podría imaginar al ver los diccionarios de datos. La relación entre las tablas, los campos llave y la cardinalidad entre éstas.

El siguente paso es explorar los datos, para esto se utilizará un nodo en clementine que permite una rápida revisión de la información como se muestra a continuación.

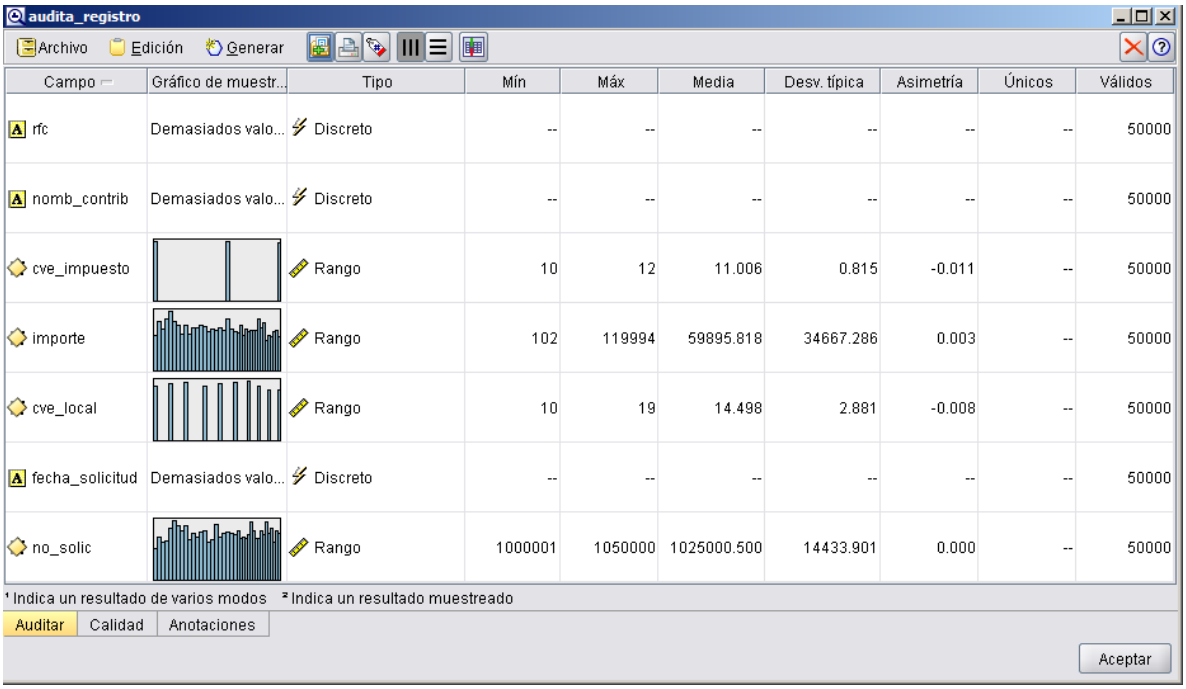


Figura III.3. Resultados de la auditoría a los campos de la tabla Registro.

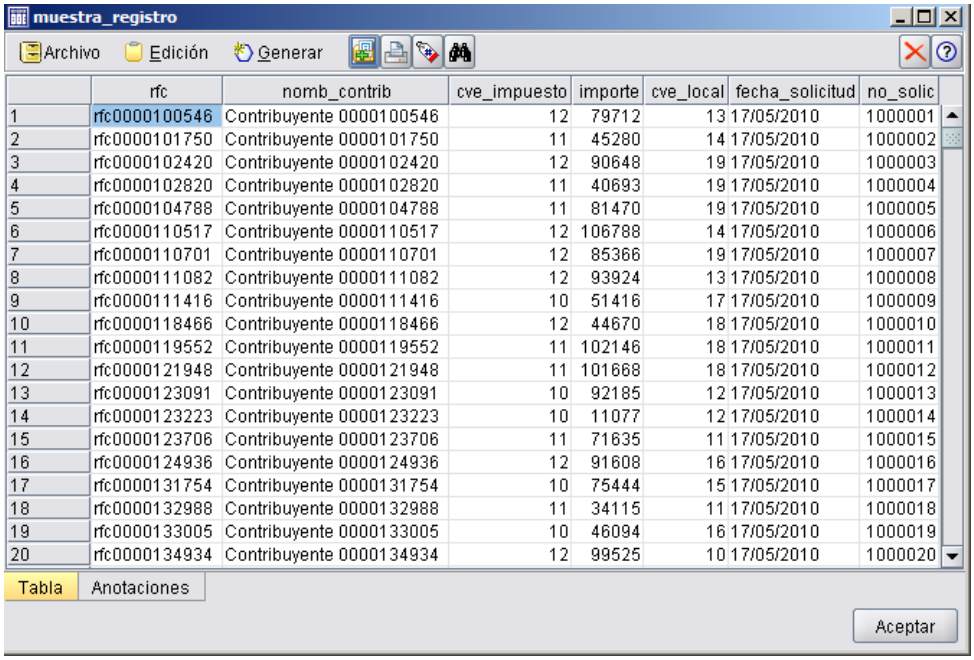
Para la tabla registro, se muestra el resultado del nodo auditar datos de Clementine (Figura III.3.), se observa el nombre del campo, un gráfico que muestra un histográma de los datos (cuando esto es posible) el tipo de dato que es, estadísticos como: mínimo, máximo, media, desviación típica, asimetría y otros datos como únicos y válidos.

Es importante observar que cuando los datos no son numéricos (como las fecha de solicitud, nombre del contribuyente), no se presenta información en los estadísticos de la auditoría.

Con esta información se puede deducir por ejemplo que la cantidad de registros de esta tabla son 50 mil solicitudes de devolución registradas.

También se puede observar que todos los valores son únicos, es decir no se repiten, que las claves de catálogos como el de la local o el impuesto son número que de entrada el minero de datos detecta como números.

Para ver una muestra de la información que aparece en la tabla se tiene lo siguiente:

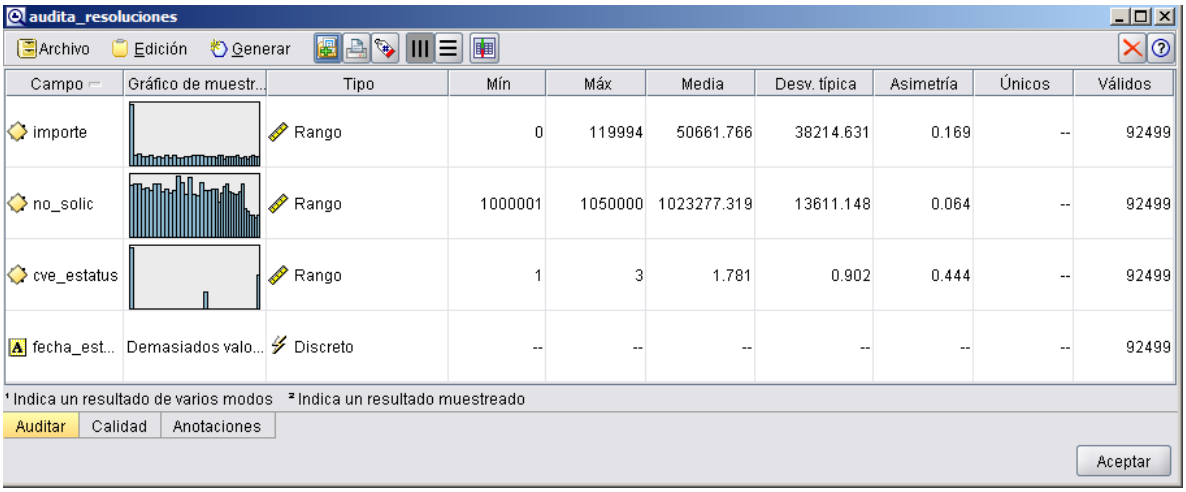


	rfc	nomb_contrib	cve_impuesto	importe	cve_local	fecha_solicitud	no_solic
1	rfc0000100546	Contribuyente 0000100546	12	79712	13	17/05/2010	1000001
2	rfc0000101750	Contribuyente 0000101750	11	45280	14	17/05/2010	1000002
3	rfc0000102420	Contribuyente 0000102420	12	90648	19	17/05/2010	1000003
4	rfc0000102820	Contribuyente 0000102820	11	40693	19	17/05/2010	1000004
5	rfc0000104788	Contribuyente 0000104788	11	81470	19	17/05/2010	1000005
6	rfc0000110517	Contribuyente 0000110517	12	106788	14	17/05/2010	1000006
7	rfc0000110701	Contribuyente 0000110701	12	85366	19	17/05/2010	1000007
8	rfc0000111082	Contribuyente 0000111082	12	93924	13	17/05/2010	1000008
9	rfc0000111416	Contribuyente 0000111416	10	51416	17	17/05/2010	1000009
10	rfc0000118466	Contribuyente 0000118466	12	44670	18	17/05/2010	1000010
11	rfc0000119552	Contribuyente 0000119552	11	102146	18	17/05/2010	1000011
12	rfc0000121948	Contribuyente 0000121948	11	101668	18	17/05/2010	1000012
13	rfc0000123091	Contribuyente 0000123091	10	92185	12	17/05/2010	1000013
14	rfc0000123223	Contribuyente 0000123223	10	11077	12	17/05/2010	1000014
15	rfc0000123706	Contribuyente 0000123706	11	71635	11	17/05/2010	1000015
16	rfc0000124936	Contribuyente 0000124936	12	91608	16	17/05/2010	1000016
17	rfc0000131754	Contribuyente 0000131754	10	75444	15	17/05/2010	1000017
18	rfc0000132988	Contribuyente 0000132988	11	34115	11	17/05/2010	1000018
19	rfc0000133005	Contribuyente 0000133005	10	46094	16	17/05/2010	1000019
20	rfc0000134934	Contribuyente 0000134934	12	99525	10	17/05/2010	1000020

Figura III.4. Muestra de información contenida en la tabla Registro.

Esto ya nos deja ver los datos en concreto como aparecen en la tabla, vemos claves de impuesto, las fechas de solicitud, en la auditoría de datos se veía muy poca información ahora se pude ver su contenido.

Para la tabla resoluciones se tiene la siguiente auditoría de datos.

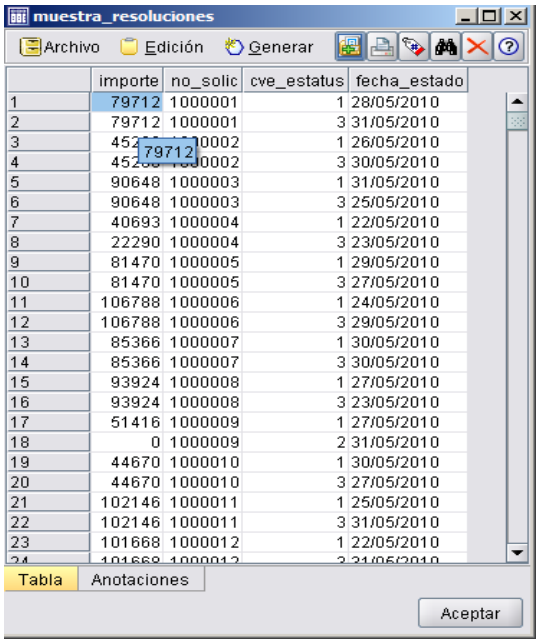


Campo	Gráfico de muestr...	Tipo	Mín	Máx	Media	Desv. típica	Asimetría	Únicos	Válidos
importe		Rango	0	119994	50661.766	38214.631	0.169	--	92499
no_solic		Rango	1000001	1050000	1023277.319	13611.148	0.064	--	92499
cve_estatus		Rango	1	3	1.781	0.902	0.444	--	92499
fecha_est...	Demasiados valo...	Discreto	--	--	--	--	--	--	92499

Figura III.5. Resultados de la auditoría a los campos de la tabla Resoluciones.

Se puede observar que la cantidad de registros es mayor en esta tabla que la de la anterior, esto era intuitivo por el diagrama entidad relación pero ahora lo podemos confirmar. Esto quiere decir que la tabla resoluciones tiene más de una solicitud de devolución.

La muestra de la información se tiene la siguiente tabla:

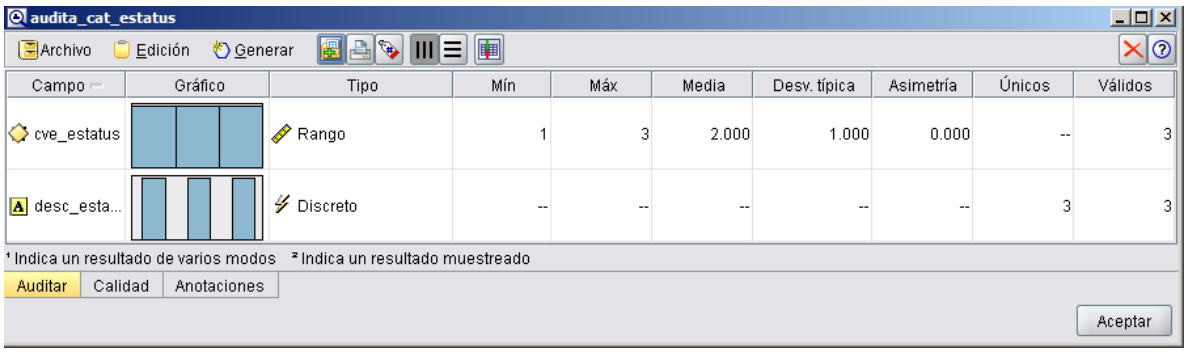


	importe	no_solic	cve_estatus	fecha_estado
1	79712	1000001	1	28/05/2010
2	79712	1000001	3	31/05/2010
3	45[79712]0		1	26/05/2010
4	45[79712]0		3	30/05/2010
5	90648	1000003	1	31/05/2010
6	90648	1000003	3	25/05/2010
7	40693	1000004	1	22/05/2010
8	22290	1000004	3	23/05/2010
9	81470	1000005	1	29/05/2010
10	81470	1000005	3	27/05/2010
11	106788	1000006	1	24/05/2010
12	106788	1000006	3	29/05/2010
13	85366	1000007	1	30/05/2010
14	85366	1000007	3	30/05/2010
15	93924	1000008	1	27/05/2010
16	93924	1000008	3	23/05/2010
17	51416	1000009	1	27/05/2010
18	0	1000009	2	31/05/2010
19	44670	1000010	1	30/05/2010
20	44670	1000010	3	27/05/2010
21	102146	1000011	1	25/05/2010
22	102146	1000011	3	31/05/2010
23	101668	1000012	1	22/05/2010

Figura III.6. Muestra de información contenida en la tabla Resoluciones.

Podemos observar que el número de solicitud se repite dependiendo de la clave de estatus en que se encuentre.

Los catálogos correspondientes a estas tablas se muestran a continuación, para el catálogo de estatus se tiene:



Campo	Gráfico	Tipo	Mín	Máx	Media	Desv. típica	Asimetría	Únicos	Válidos
cve_estatus		Rango	1	3	2.000	1.000	0.000	--	3
desc_esta...		Discreto	--	--	--	--	--	3	3

Figura III.7. Resultados de la auditoría a los campos de la tabla catálogo de estatus.

Se observa que este catálogo tiene 3 posibles valores, para ver su contenido se tiene la siguiente tabla.

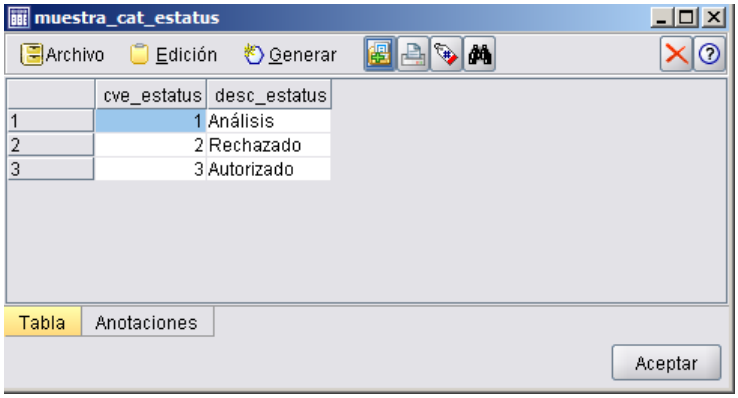


Figura III.8. Información contenida en la tabla catálogo de estatus.

Con esta información ya sabemos que existen 3 posibles estados en los que se encuentra una solicitud de devolución.

Para los impuestos se tiene el siguiente reporte:

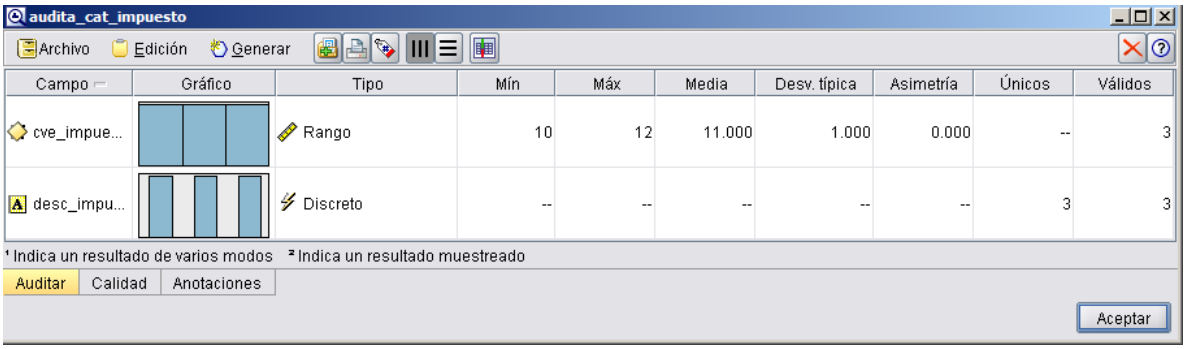


Figura III.9. Resultados de la auditoría a los campos de la tabla catálogo de impuesto.

Observando los valores en la tabla se tiene:

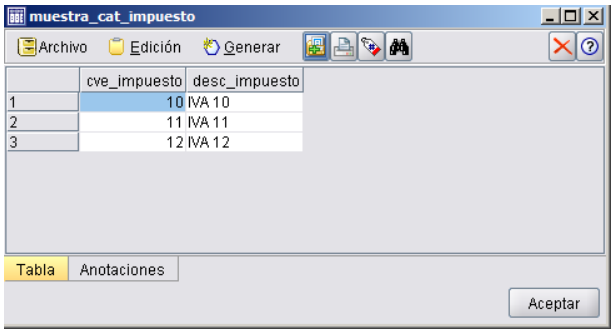


Figura III.10. Información contenida en la tabla catálogo de impuesto.

Existen 3 tipos de IVA para las solicitudes que se clasifican con números de 10 a 12.

Por último la tabla de catálogo de Administraciones Locales de Recaudación:

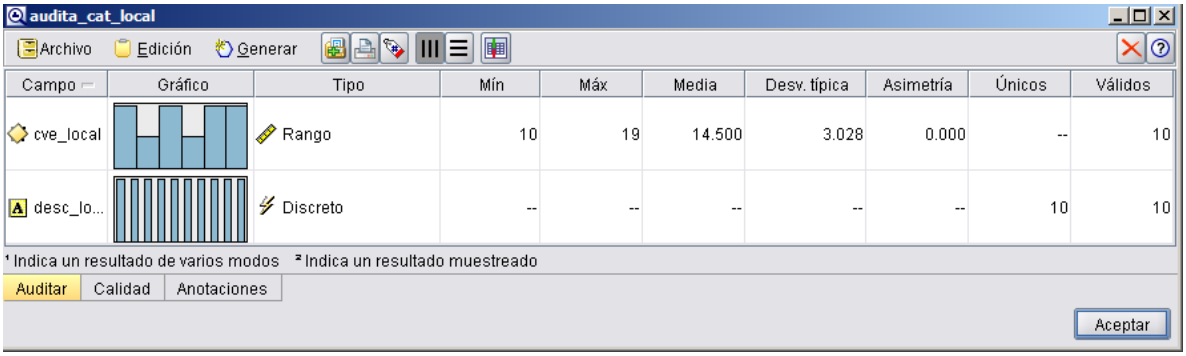


Figura III.11. Resultados de la auditoría a los campos de la tabla catálogo de local.

El contenido de esta tabla es:

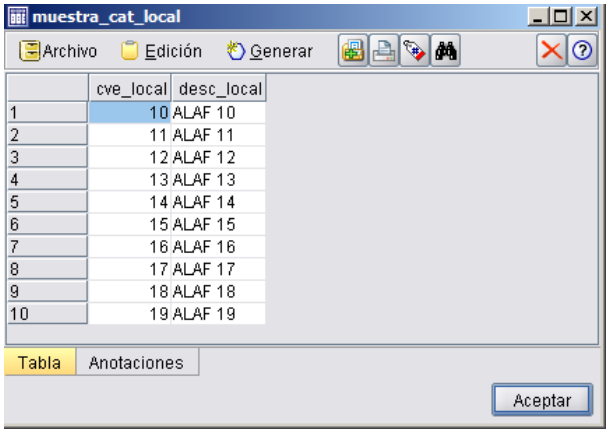


Figura III.12. Muestra de información contenida en la tabla catálogo de local.

Se puede observar que los 3 catálogos tienen números para identificar cada uno de los valores, el valor nominal no tiene sentido ya que solo se utiliza como identificador, es decir, la local 10 no significa que sea menor que la 16, simplemente es una etiqueta para facilitar el manejo de la información.

La ruta en Clementine que muestra esta información se presenta a continuación:

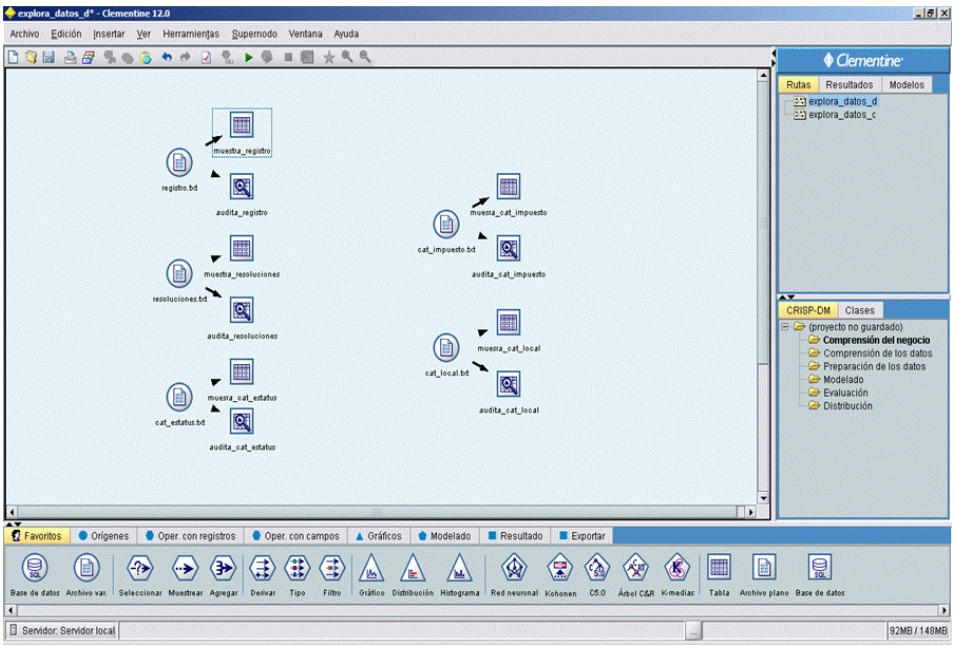


Figura III.13. Ruta en Clementine que genera las auditorías y las muestras de la información para devoluciones.

En primer lugar, las fuentes de datos son archivos txt, estos tienen conectados dos nodos de resultados: uno de auditoría de datos y otro que los muestra a través de una tabla.

III.1.2. Cobranza.

Para la cobranza se tendrían las tablas donde se tiene la información del registro, etapas de avance en la cobranza y las bajas de los créditos fiscales.

La definición de las tablas que guardan la información de los créditos fiscales sería la siguiente:

Registro de adeudos: Registro_adeudo		
Nombre	Definición	Descripción
rfc	char(13)	Registro federal de contribuyentes
nomb_contrib	char(255)	Nombre del contribuyente
no_credito	int(10)	Número de crédito fiscal
cve_impuesto	int(2)	Clave de impuesto origen del crédito fiscal
cve_local	int(2)	Clave de local de recaudación

importe	real(12,2)	Importe del crédito fiscal
fecha_origen	Date	Fecha de origen del crédito fiscal
fecha_registro	Date	Fecha de registro

Tabla III.6. Definición tabla registro de adeudos.

La tabla de registro de adeudos donde se tienen todos los créditos fiscales actuales e históricos

Etapas de cobro: Etapas		
Nombre	Definición	Descripción
no_credito	int(10)	Número de crédito fiscal
cve_etapa	int(2)	Clave de etapa de cobro
fecha_etapa	date	Fecha de la etapa del cobro

Tabla III.7. Definición tabla etapas de cobro.

Una tabla donde se registra la etapa de cobro en que se encuentra el crédito fiscal.

Bajas de créditos: Bajas		
Nombre	Definición	Descripción
no_credito	int(10)	Número de crédito fiscal
cve_baja	int(2)	Clave de baja
fecha_baja	date	Fecha de baja

Tabla III.8. Definición tabla baja de créditos.

La tabla registra el fin del crédito fiscal y el resultado por el cual se generó la baja.

Catálogo de impuestos: Cat_impuesto		
Nombre	Definición	Descripción
cve_impuesto	int(2)	Clave de impuesto del crédito fiscal
desc_impuesto	char(50)	Descripción del impuesto

Tabla III.9. Definición tabla catálogo de impuestos.

El catálogo de impuestos por los que se origina el crédito fiscal.

Catálogo de etapas de cobro: Cat_etapa		
Nombre	Definición	Descripción
cve_etapa	int(2)	Clave de etapa de cobro
desc_etapa	char(50)	Descripción de etapa

Tabla III.10. Definición tabla catálogo de etapas de cobro.

La tabla catálogo de etapas de cobro tiene la descripción en la cual se encuentra actualmente el crédito fiscal.

Catálogo de locales: Cat_local		
Nombre	Definición	Descripción
cve_local	int(2)	Clave de local de recaudación
desc_local	char(50)	Descripción de local

Tabla III.11. Definición tabla catálogo de locales.

Una tabla con la descripción de la local de recaudación que controla el crédito fiscal.

Catálogo de claves de baja: Cat_baja		
Nombre	Definición	Descripción
cve_baja	int(2)	Clave de baja
desc_baja	char(50)	Descripción de baja

Tabla III.12. Definición tabla catálogo de claves de baja.

El catálogo de las claves de baja finales para cada crédito fiscal.

El diagrama entidad – relación para estas tablas es el siguiente:

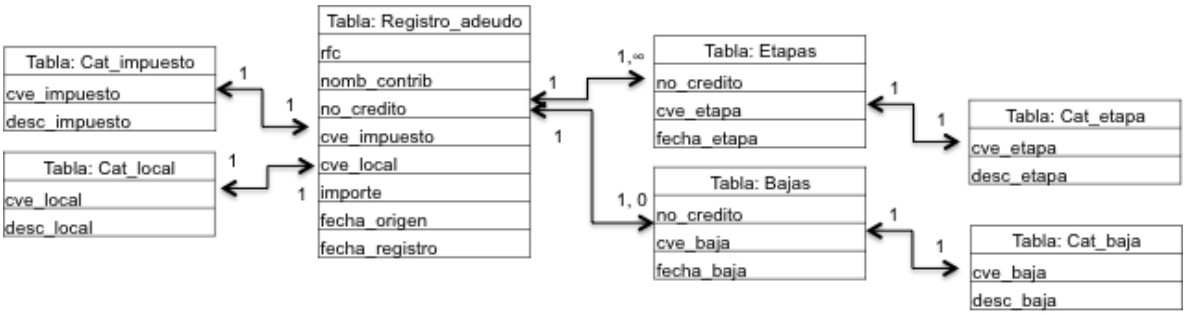


Figura III.14. Diagrama entidad relación para las tablas de cobranza.

La tabla etapas puede tener màs de un registro por etapa de cobro para los créditos fiscales, la tabla bajas tendrá uno o ningún registro dependiendo si el crédito fiscal es activo.

Para ver la información de estas tablas se muestra primeramente la auditoría de los campos para la tabla registro_adeudo.

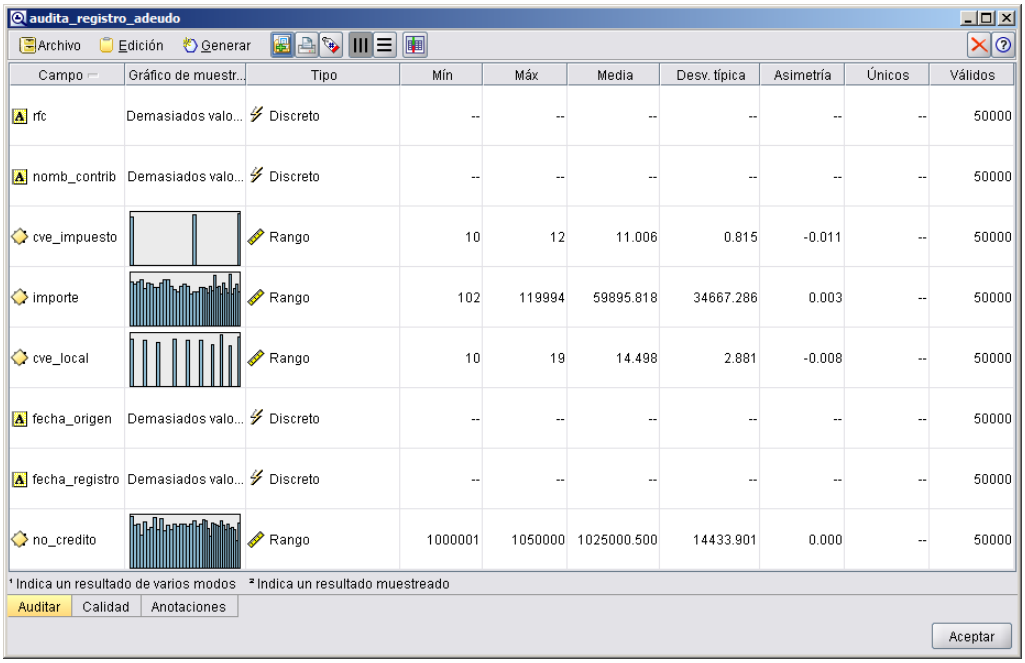
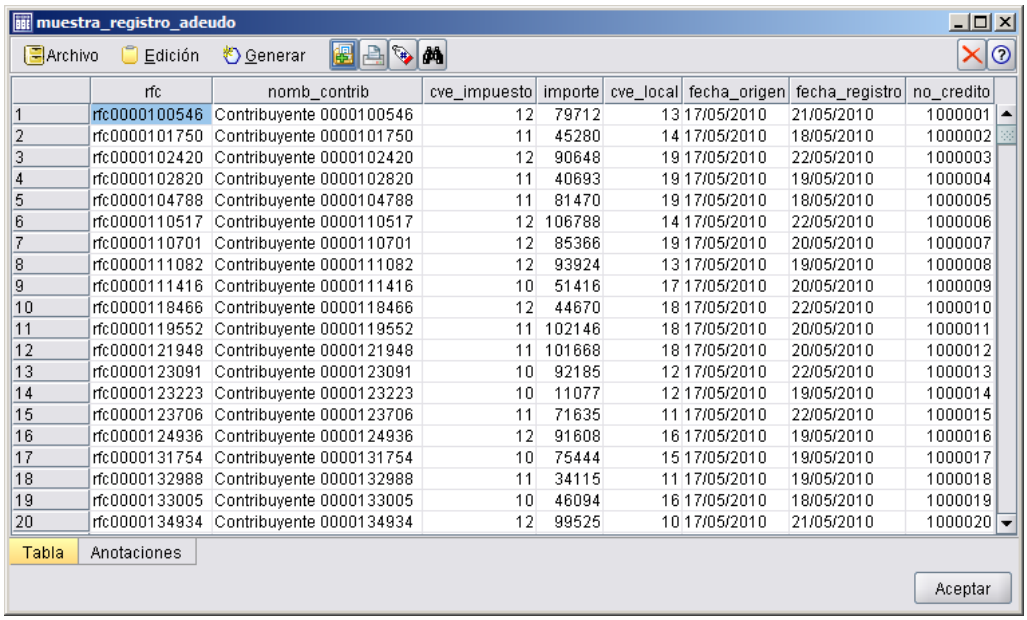


Figura III.15. Auditoría de los campos para la tabla registro_adeudo.

Se pueden observar 50 mil registros (Válidos) que contienen créditos fiscales junto con los demás campos de la tabla.

Una muestra de esta información es la que sigue:



muestra_registro_adeudo

Archivo Edición Generar

	rfc	nomb_contrib	cve_impuesto	importe	cve_local	fecha_origen	fecha_registro	no_credito
1	rfc0000100546	Contribuyente 0000100546	12	79712	13	17/05/2010	21/05/2010	1000001
2	rfc0000101750	Contribuyente 0000101750	11	45280	14	17/05/2010	18/05/2010	1000002
3	rfc0000102420	Contribuyente 0000102420	12	90648	19	17/05/2010	22/05/2010	1000003
4	rfc0000102820	Contribuyente 0000102820	11	40693	19	17/05/2010	19/05/2010	1000004
5	rfc0000104788	Contribuyente 0000104788	11	81470	19	17/05/2010	18/05/2010	1000005
6	rfc0000110517	Contribuyente 0000110517	12	106788	14	17/05/2010	22/05/2010	1000006
7	rfc0000110701	Contribuyente 0000110701	12	85366	19	17/05/2010	20/05/2010	1000007
8	rfc0000111082	Contribuyente 0000111082	12	93924	13	17/05/2010	19/05/2010	1000008
9	rfc0000111416	Contribuyente 0000111416	10	51416	17	17/05/2010	20/05/2010	1000009
10	rfc0000118466	Contribuyente 0000118466	12	44670	18	17/05/2010	22/05/2010	1000010
11	rfc0000119552	Contribuyente 0000119552	11	102146	18	17/05/2010	20/05/2010	1000011
12	rfc0000121948	Contribuyente 0000121948	11	101668	18	17/05/2010	20/05/2010	1000012
13	rfc0000123091	Contribuyente 0000123091	10	92185	12	17/05/2010	22/05/2010	1000013
14	rfc0000123223	Contribuyente 0000123223	10	11077	12	17/05/2010	19/05/2010	1000014
15	rfc0000123706	Contribuyente 0000123706	11	71635	11	17/05/2010	22/05/2010	1000015
16	rfc0000124936	Contribuyente 0000124936	12	91608	16	17/05/2010	19/05/2010	1000016
17	rfc0000131754	Contribuyente 0000131754	10	75444	15	17/05/2010	19/05/2010	1000017
18	rfc0000132988	Contribuyente 0000132988	11	34115	11	17/05/2010	19/05/2010	1000018
19	rfc0000133005	Contribuyente 0000133005	10	46094	16	17/05/2010	18/05/2010	1000019
20	rfc0000134934	Contribuyente 0000134934	12	99525	10	17/05/2010	21/05/2010	1000020

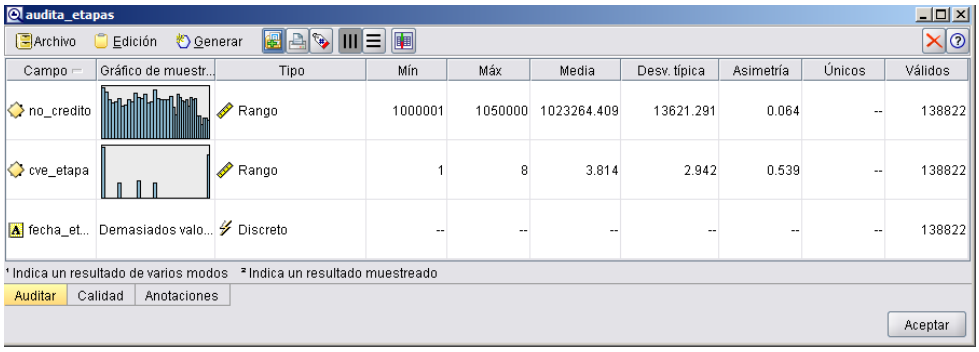
Tabla Anotaciones

Aceptar

Figura III.16. Muestra de información contenida en la tabla registro_adeudo.

Se puede observar que un crédito fiscal le corresponde a un contribuyente con el importe y la clave de impuesto, generada por una ALR (cve_local) con las fechas en que se originó y se registró.

La siguiente tabla a explorar es Etapas:



audita_etapas

Archivo Edición Generar

Campo	Gráfico de muestr...	Tipo	Mín	Máx	Media	Desv. típica	Asimetría	Únicos	Válidos
no_credito		Rango	1000001	1050000	1023264.409	13621.291	0.064	--	138822
cve_etapa		Rango	1	8	3.814	2.942	0.539	--	138822
fecha_et...	Demasiados valo...	Discreto	--	--	--	--	--	--	138822

[1] Indica un resultado de varios modos [2] Indica un resultado muestreado

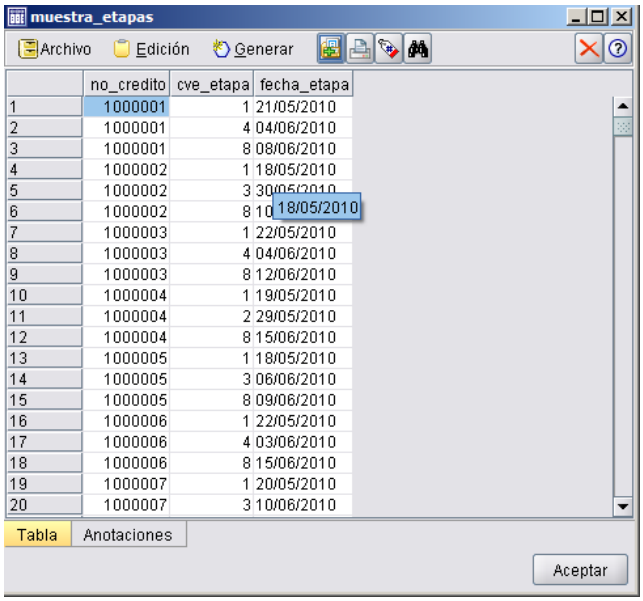
Auditar Calidad Anotaciones

Aceptar

Figura III.17. Auditoría de los campos para la tabla etapas.

En la columna Validos hay 138,822 registros, lo que representa màs de 3 veces la cantidad que tiene la tabla registro_adeudos, esto nos puede hacer pensar que hay más de una etapa por crédito.

Para ver la información se observa la muestra de esta tabla:

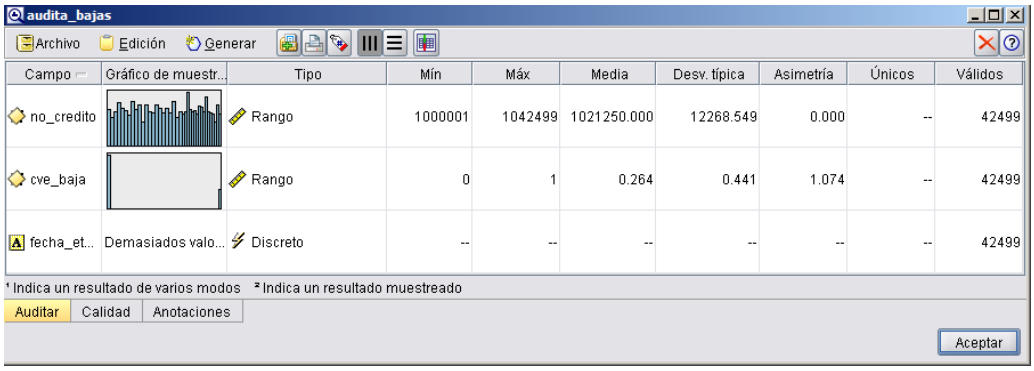


	no_credito	cve_etapa	fecha_etapa
1	1000001	1	21/05/2010
2	1000001	4	04/06/2010
3	1000001	8	08/06/2010
4	1000002	1	18/05/2010
5	1000002	3	30/05/2010
6	1000002	8	10/06/2010
7	1000003	1	22/05/2010
8	1000003	4	04/06/2010
9	1000003	8	12/06/2010
10	1000004	1	19/05/2010
11	1000004	2	29/05/2010
12	1000004	8	15/06/2010
13	1000005	1	18/05/2010
14	1000005	3	06/06/2010
15	1000005	8	09/06/2010
16	1000006	1	22/05/2010
17	1000006	4	03/06/2010
18	1000006	8	15/06/2010
19	1000007	1	20/05/2010
20	1000007	3	10/06/2010

Figura III.18. Muestra de información contenida en la tabla etapas.

Con esta exploración podemos confirmar que existen varias etapas para un crédito fiscal que se registran en diferentes momentos.

Para la tabla bajas se tiene la siguiente información:



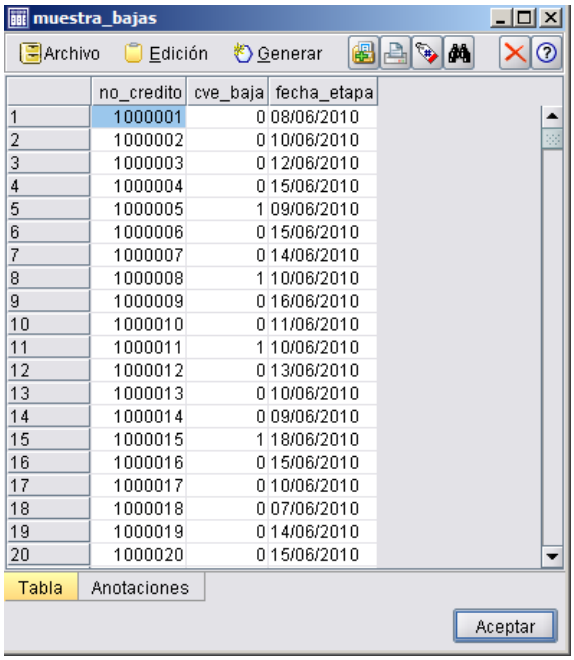
Campo	Gráfico de muestr...	Tipo	Mín	Máx	Media	Desv. típica	Asimetría	Únicos	Válidos
no_credito		Rango	1000001	1042499	1021250.000	12268.549	0.000	--	42499
cve_baja		Rango	0	1	0.264	0.441	1.074	--	42499
fecha_et...	Demasiados valo...	Discreto	--	--	--	--	--	--	42499

[1] Indica un resultado de varios modos [2] Indica un resultado muestreado

Figura III.19. Auditoría de los campos para la tabla bajas.

Se puede observar que los registros válidos son menores a la de registro de adeudos, esto es porque no todos han terminado en una baja, es decir existen adeudos activos.

En la tabla con los registros se puede ver la información:

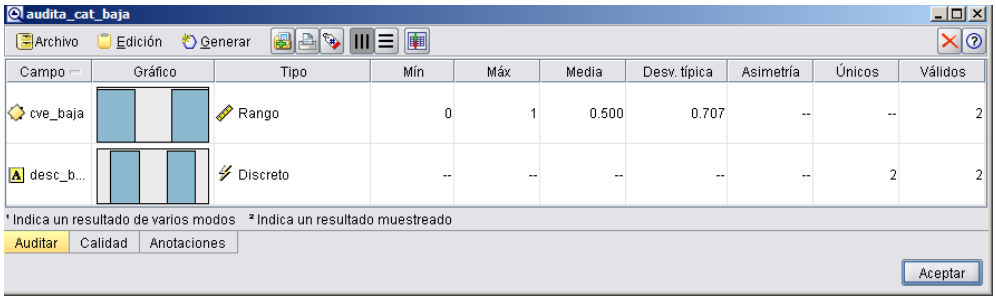


	no_credito	cve_baja	fecha_etapa
1	1000001	0	08/06/2010
2	1000002	0	10/06/2010
3	1000003	0	12/06/2010
4	1000004	0	15/06/2010
5	1000005	1	09/06/2010
6	1000006	0	15/06/2010
7	1000007	0	14/06/2010
8	1000008	1	10/06/2010
9	1000009	0	16/06/2010
10	1000010	0	11/06/2010
11	1000011	1	10/06/2010
12	1000012	0	13/06/2010
13	1000013	0	10/06/2010
14	1000014	0	09/06/2010
15	1000015	1	18/06/2010
16	1000016	0	15/06/2010
17	1000017	0	10/06/2010
18	1000018	0	07/06/2010
19	1000019	0	14/06/2010
20	1000020	0	15/06/2010

Figura III.20. Muestra de información contenida en la tabla bajas.

Se puede observar que cada crédito fiscal tiene asignada una clave de baja, como veremos después en el catálogo de esta información corresponde al tipo de baja en que puede terminar un crèdito.

Al auditar el catálogo de las claves de baja se tiene lo siguiente:



Campo	Gráfico	Tipo	Mín	Máx	Media	Desv. típica	Asimetría	Únicos	Válidos
cve_baja		Rango	0	1	0.500	0.707	--	--	2
desc_b...		Discreto	--	--	--	--	--	2	2

[1] Indica un resultado de varios modos [2] Indica un resultado muestreado

Figura III.21. Auditoría de los campos para la tabla cat_baja.

Esto corresponde a la siguiente información de la tabla:

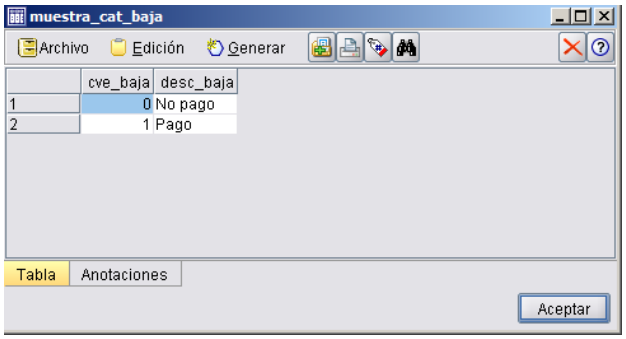


Figura III.22. Información contenida en la tabla bajas.

Aquí ya se puede observar la clave que corresponde al pago o no pago de un crédito fiscal.

Para el catálogo de etapas se tiene la siguente auditoría:

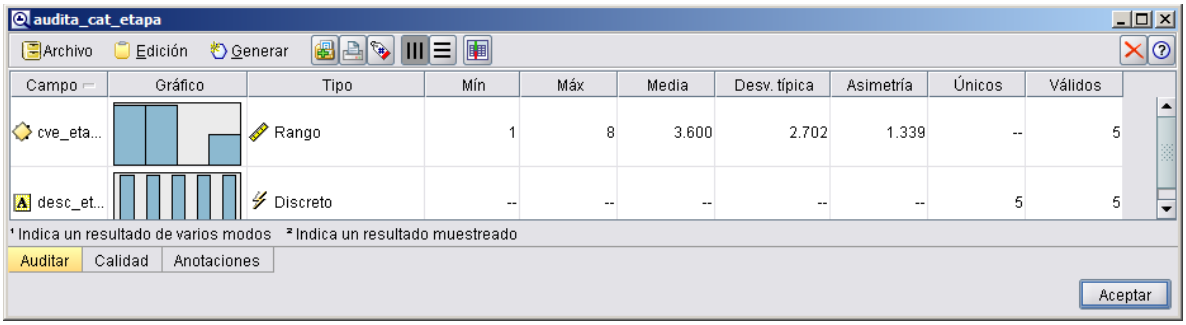


Figura III.23. Auditoría de los campos para la tabla cat_etapa.

Para tener una mejor idea de los datos es más útil revisar el contenido:

Figura III.24. Información contenida en la tabla cat_etapa.

Se puede observar que se tienen 5 etapas que corresponden al ciclo de la cobranza que inicia con una notificación y finaliza con una baja.

Para el catálogo de impuestos se tiene:

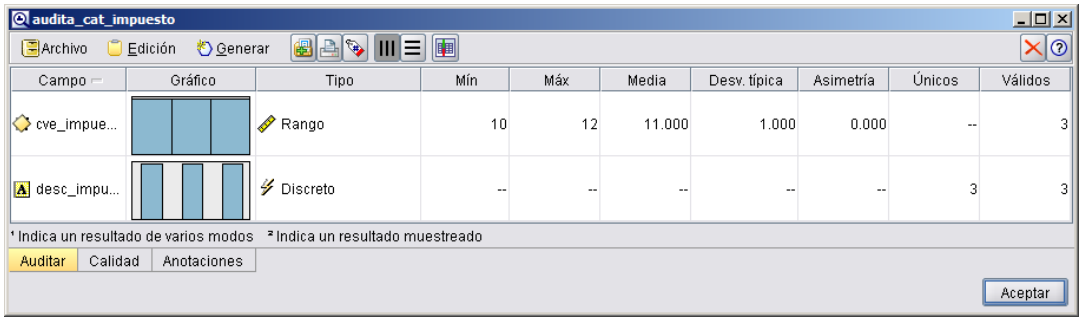


Figura III.25. Auditoría de los campos para la tabla cat_impuesto.

La información de la tabla es la siguiente:

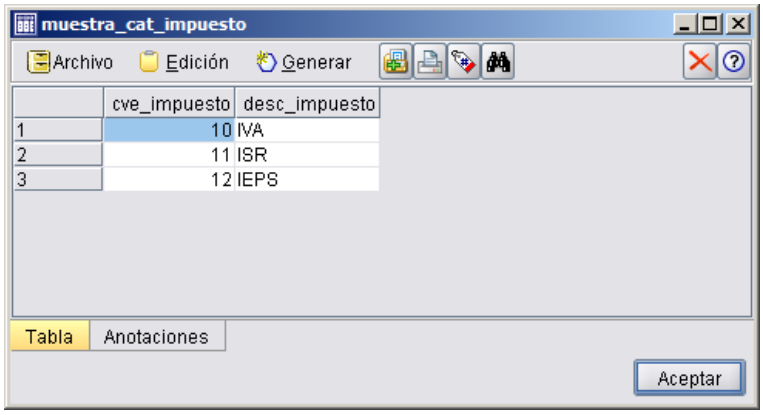


Figura III.26. Información contenida en la tabla cat_impuesto.

Con lo que se ha podido observar los catálogos muestran información importante de las diferentes características en este caso de los crétidos fiscales, con esto se tiene conocimiento que existen 3 tipos de impuestos para la cobranza.

Por último el catálogo de locales:

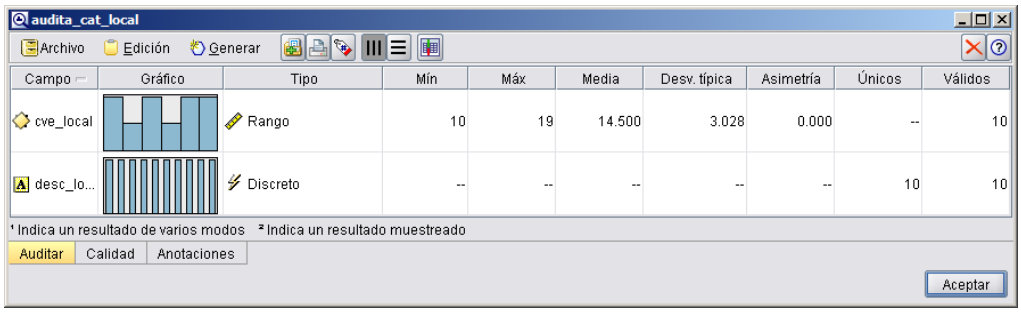


Figura III.27. Auditoría de los campos para la tabla cat_local.

Muestra la siguiente información:

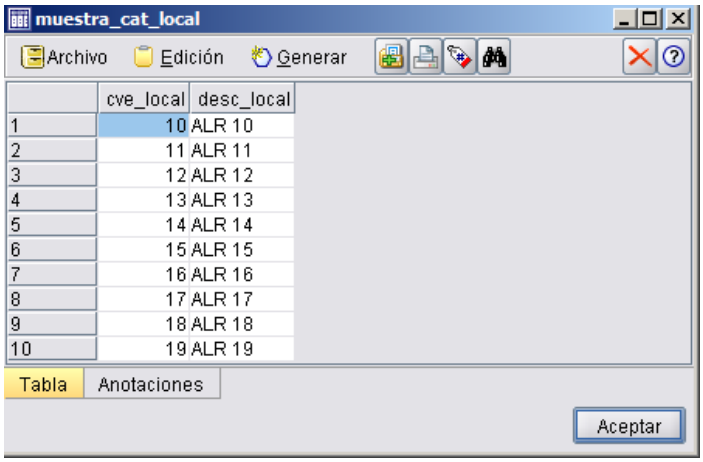


Figura III.28. Información contenida en la tabla cat_local.

Corresponde a las 10 ALR que controlan la cobranza de los créditos fiscales.

La ruta en Clementine que se utilizó para obtener esta información se muestra a continuación:

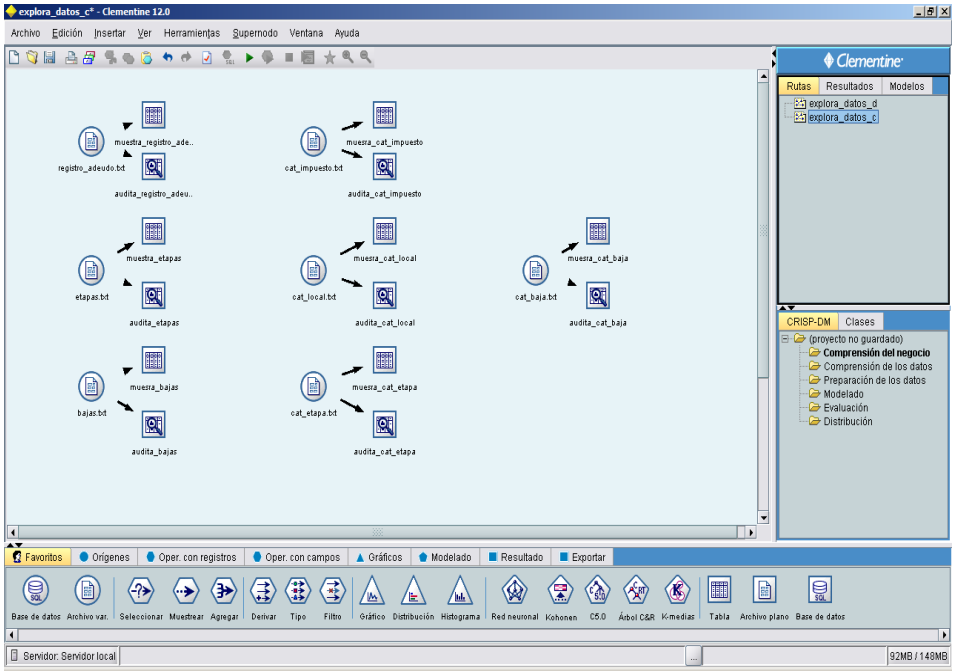


Figura III.29. Ruta en Clementine que genera las auditorías y las muestras de la información para cobranza.

III.2. Preparación de los datos.

Una vez que se tiene un entendimiento de los datos se debe pasar a la preparación de los mismos, no olvidemos que el fin de estas etapas es crear la vista minable que se utilizará para la generación de los modelos.

Se tiene que empezar por juntar los datos en una tabla principal, de manera que se puedan tener las variables independientes y dependientes.

Una parte fundamental en esta etapa es definr la variable objetivo (dependiente), ésta será la que se utilizará para encontrar la respuesta al problema que se desea resolver, para el caso del presente trabajo será encontrar la variable resolución para las devoluciones y pago para la cobranza.

III.2.1. Devoluciones.

Como ya se observó en el entendimiento de los datos, en la tabla registro se tiene la entrada de las solcitudes de devolución, sin embargo en estos campos no se tiene aún conocimiento de qué fue lo que pasó con las solicitudes.

En la tabla de resoluciones se tiene el campo cve_estatus, como ya se vió en la exploración de los datos para su catálogo, se tienen 3 posibles estatus que son: Análisis, rechazado y autorizado.

Para obtener la varialbe objetivo se va a utilizar los valores de estos campos, el primero corresponde a las solicitudes que aún no se resuelven, los últimos dos son las resueltas, para este último grupo se pueden tener dos valores: rechazado o autorizado.

Estos son los valores de nuestra variable objetivo, son los valores conocidos, con los que el modelo se entrenará para dar un pronóstico sobre los que aún no se resuelven (en Análisis). Para construir la vista minable se puede empezar por la preparación de esta variable.

La vista minable será aquella que cuente con la historia, entonces se deben filtrar los registros de la tabla resoluciones por aquellos que son diferentes a análisis: cve_estatus diferente a 1, así nos quedaremos solo con los valres 2 o 3, autorizado y rechazado respectívamente.

Una vez que se tenga esta información entonces la variable objetivo se puede reclasificar en una variable que le llamaremos resolución, la cual tendrá el valor R para los rechazados y el valor A para los autorizados.

De esta tabla también sería importante obtener ei importe, éste es lo que se le devolvió al contribuyente y la fecha de estado, esta última para saber el tiempo que transcurrió en el trámite de resolución de la devolución.

La ruta que realiza este proceso en Clementine es la siguiente:

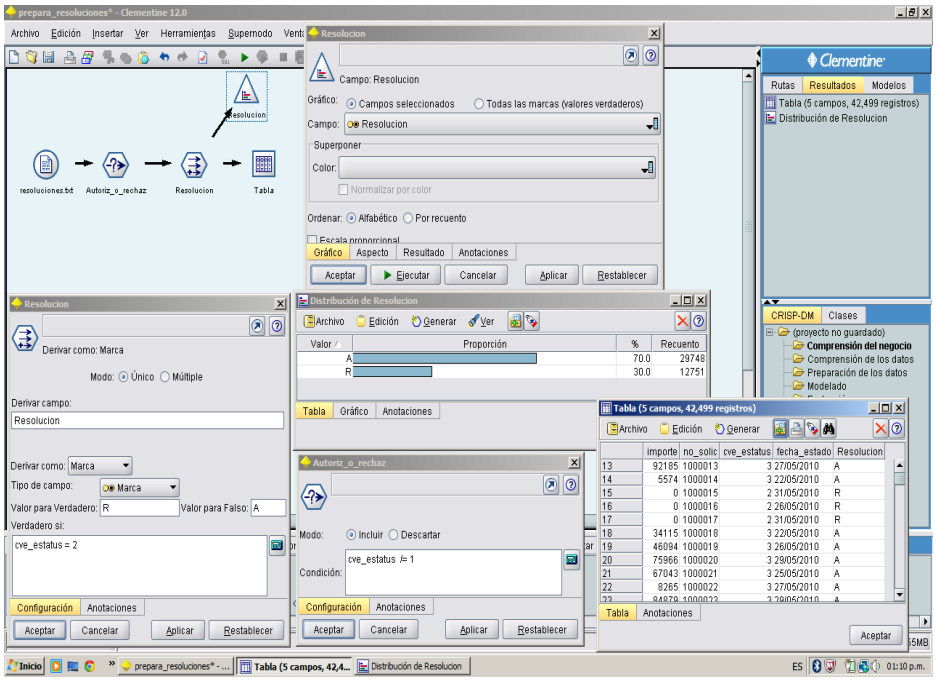


Figura III.30. Ruta en Clementine que muestra el inicio de preparación de los datos para las resoluciones.

Al nodo de las resoluciones se le añade una selección que quita el estatus =1, después se crea la variable resolucion con un nodo derivar definiendo el valor R para Rechazado y A para Autorizado.

Se puede ver el resultado en la tabla de salida, para la frecuencia de los valores se utiliza un nodo de graficado llamado Distribución, éste presenta el recuento de los casos de la variable objetivo, se puede observar que el 70% de las resoluciones se autorizan mientras que el 30 % se rechaza.

Esta información se adjuntará a la que se tiene en la tabla Registros, con esto se tendrá el ciclo de las devoluciones (el registro y la resolución) en una sola tabla.

Esto se realiza con el nodo Fundir en Clementine, la llave para poder unir estas dos tablas es el campo es el número de solicitud (no_solicitud). Recordando el diagrama de entidad relación este es el campo llave entre las dos tablas.

Es importante mencionar que no solo el campo llave se llama igual entre las dos tablas, sino también el campo importe, sin embargo no tiene los mismos valores, de acuerdo el diccionario de datos el de la tabla Registro es el que el contribuyente solicitó y el de la tabla resoluciones es el que el dictaminador resolivió.

Cuando se unan las dos tablas solo uno de los campos permanecerá en el archivo, para resolver este inconveniente se puede cambiar el nombre de cada uno por: importe_solicitado e importe_autorizado.

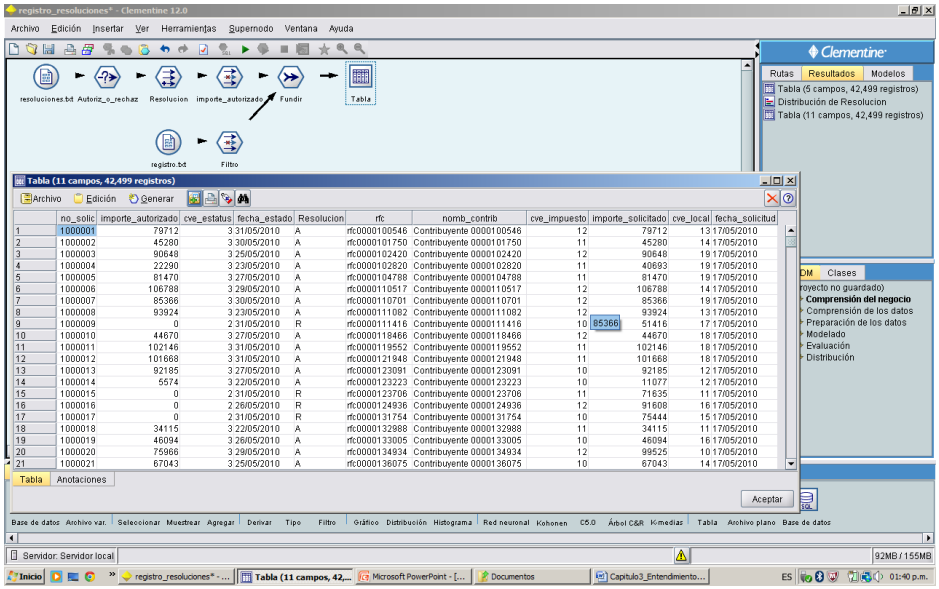
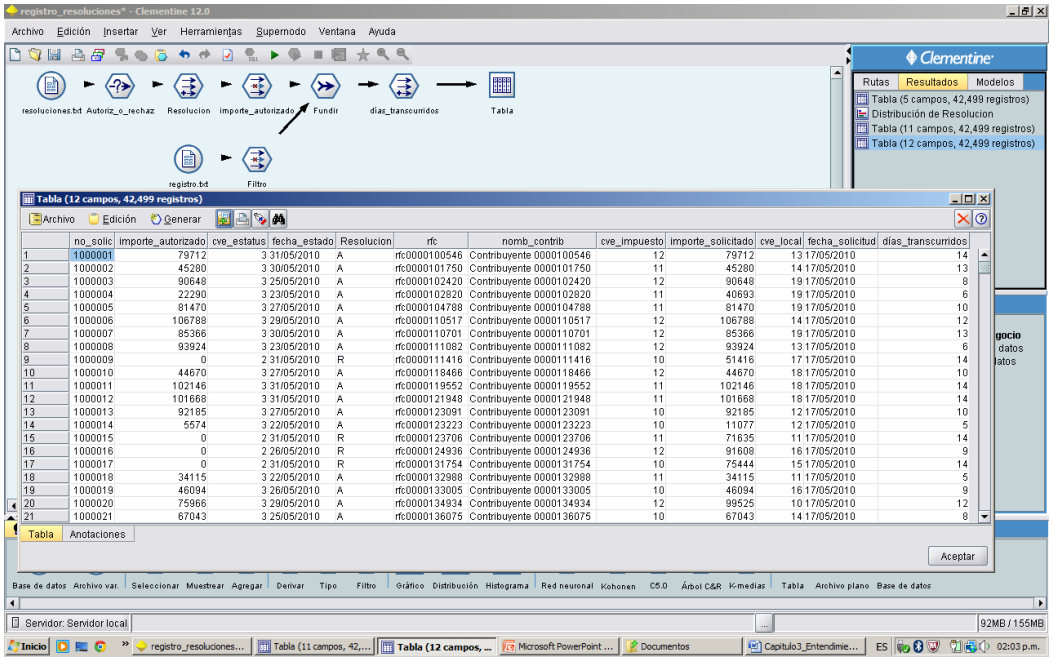


Figura III.31. Ruta en Clementine que muestra el cruce entre las resoluciones y el registro para devoluciones.

Con estos pasos se tiene en principio una vista minable, ya que se tiene la variable objetivo y posibles variables de entrada, sin embargo aún se pueden construir variables que podrían ser interesantes explorar para ser consideradas en un modelo.

Para este caso, se va a construir una variable que tendrá el número de días que transcurrieron desde que se generó la solicitud hasta su resolución.

Para esto se necesita realizar un cálculo entre la fecha de solicitud y fecha de estado para conocer la diferencia entre ambas.



	no_solic	importe_autorizado	cve_estatus	fecha_estado	Resolucion	rfc	nomb_contrib	cve_impuesto	importe_solicitado	cve_local	fecha_solicitud	dias_transcurridos
1	1000001	79712	3	31/05/2010	A	rfc0000100546	Contribuyente 0000100546	12	79712	13	17/05/2010	14
2	1000002	45280	3	30/05/2010	A	rfc0000101750	Contribuyente 0000101750	11	45280	14	17/05/2010	13
3	1000003	90648	3	25/05/2010	A	rfc0000102420	Contribuyente 0000102420	12	90648	19	17/05/2010	8
4	1000004	22290	3	23/05/2010	A	rfc0000102820	Contribuyente 0000102820	11	40693	19	17/05/2010	6
5	1000005	81470	3	27/05/2010	A	rfc0000104788	Contribuyente 0000104788	11	81470	19	17/05/2010	10
6	1000006	106788	3	29/05/2010	A	rfc0000110517	Contribuyente 0000110517	12	106788	14	17/05/2010	12
7	1000007	85366	3	30/05/2010	A	rfc0000110701	Contribuyente 0000110701	12	85366	19	17/05/2010	13
8	1000008	93924	3	23/05/2010	A	rfc0000111082	Contribuyente 0000111082	12	93924	13	17/05/2010	6
9	1000009	0	2	31/05/2010	R	rfc0000111416	Contribuyente 0000111416	10	51416	17	17/05/2010	14
10	1000010	44670	3	27/05/2010	A	rfc0000118466	Contribuyente 0000118466	12	44670	18	17/05/2010	10
11	1000011	102146	3	31/05/2010	A	rfc0000119552	Contribuyente 0000119552	11	102146	18	17/05/2010	14
12	1000012	101668	3	31/05/2010	A	rfc0000121948	Contribuyente 0000121948	11	101668	18	17/05/2010	14
13	1000013	92185	3	27/05/2010	A	rfc0000123091	Contribuyente 0000123091	10	92185	12	17/05/2010	10
14	1000014	5574	3	22/05/2010	A	rfc0000123223	Contribuyente 0000123223	10	11077	12	17/05/2010	5
15	1000015	0	2	31/05/2010	R	rfc0000123706	Contribuyente 0000123706	11	71635	11	17/05/2010	14
16	1000016	0	2	26/05/2010	R	rfc0000124936	Contribuyente 0000124936	12	91608	16	17/05/2010	9
17	1000017	0	2	31/05/2010	R	rfc0000131754	Contribuyente 0000131754	10	75444	15	17/05/2010	14
18	1000018	34115	3	22/05/2010	A	rfc0000132988	Contribuyente 0000132988	11	34115	11	17/05/2010	5
19	1000019	46094	3	26/05/2010	A	rfc0000133005	Contribuyente 0000133005	10	46094	16	17/05/2010	9
20	1000020	75966	3	29/05/2010	A	rfc0000134934	Contribuyente 0000134934	12	99525	10	17/05/2010	12
21	1000021	67043	3	25/05/2010	A	rfc0000136075	Contribuyente 0000136075	10	67043	14	17/05/2010	8

Figura III.32. Muestra el cruce entre las resoluciones y el registro para devoluciones y campo adicional creado.

Con un nodo derivar se puede crear la nueva variable la cual tendrá el tiempo en días que tardó en resolverse una solicitud de devolución.

A esta información para tenerla completa se necesitaría adjuntarle las descripciones del impuesto y de la local de sus respectivos catálogos.

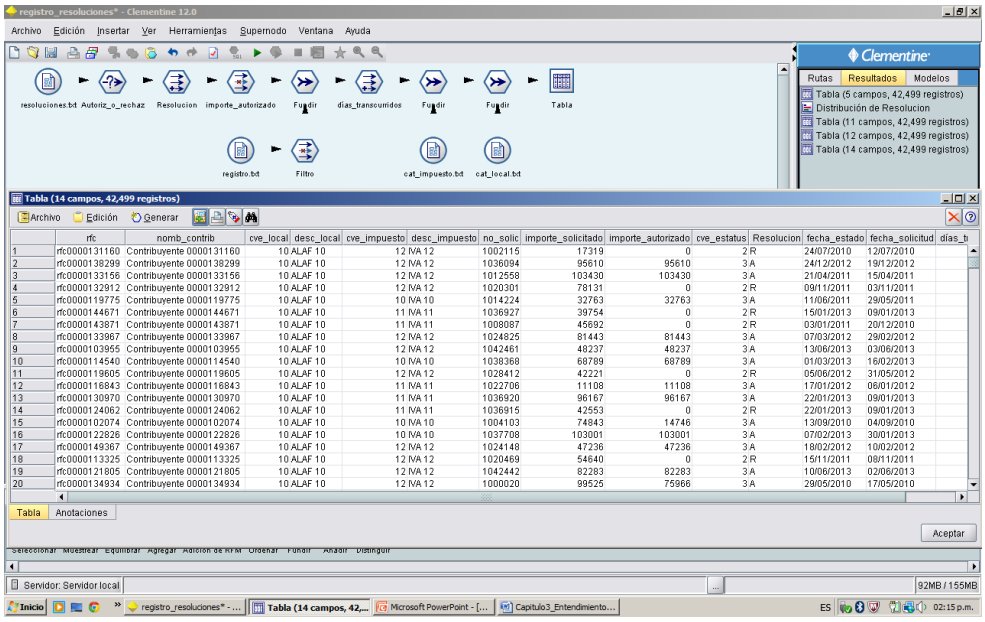


Figura III.33. Muestra el cruce entre las resoluciones,el registro y los catálogos de impuesto y local.

Ordenando la información se tiene una primera aproximación a la vista minable que se utilizará en la parte del modelado.

Con esta información se puede empezar a analizar el comportamiento de las variables, vemos por ejemplo la distribución de las ALAF.

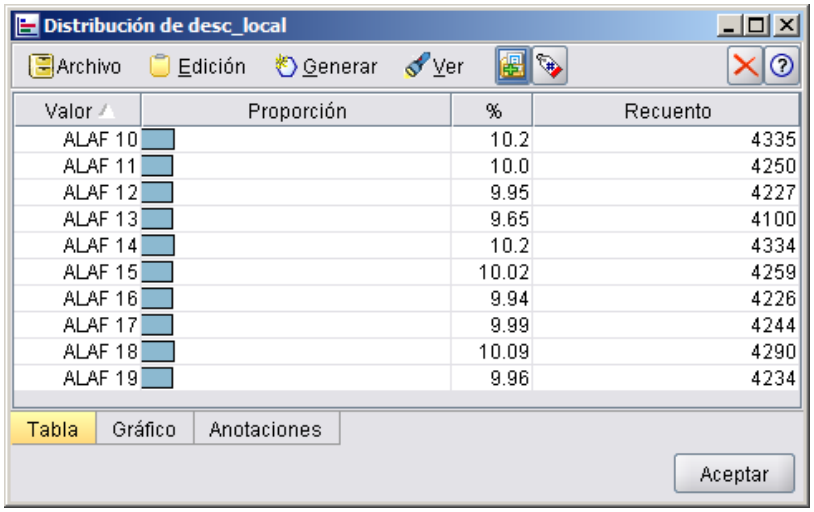


Figura III.34. Distribución de la variable desc_local.

Para el tipo de impuesto se tiene la siguiente distribución:

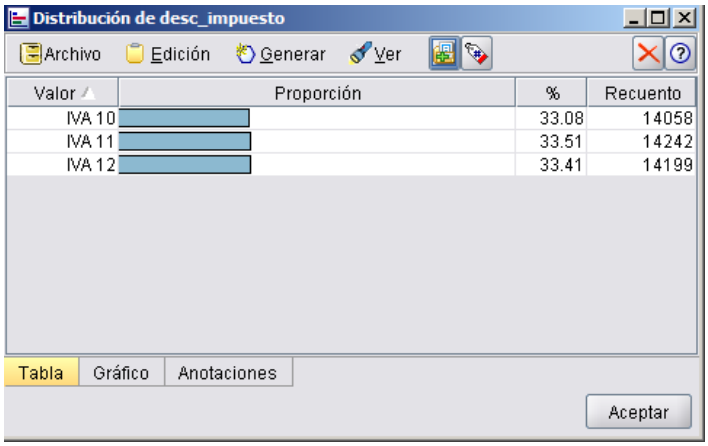


Figura III.35. Distribución de la variable desc_impuesto.

El importe solicitado tiene el siguiente histograma:

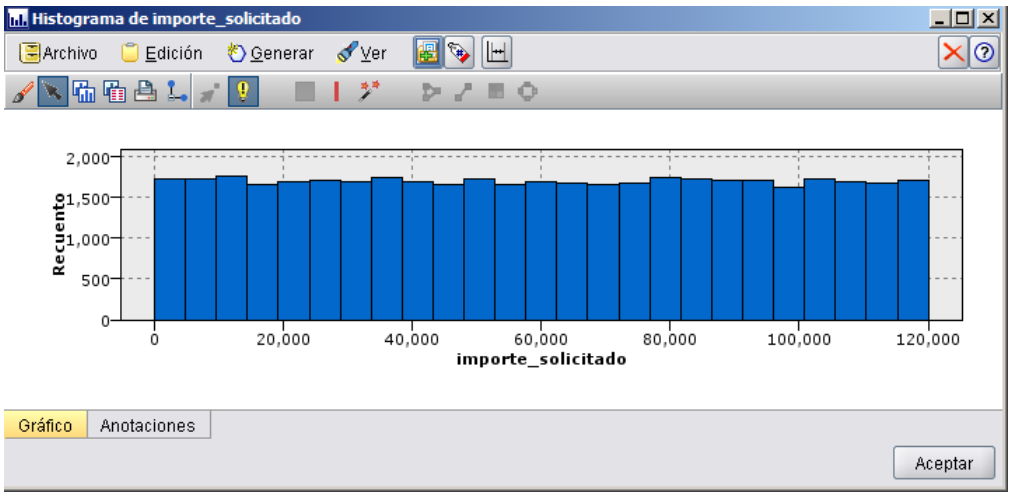


Figura III.36. Histograma de la variable importe_solicitado.

El histograma de los días transcurridos es el siguiente:

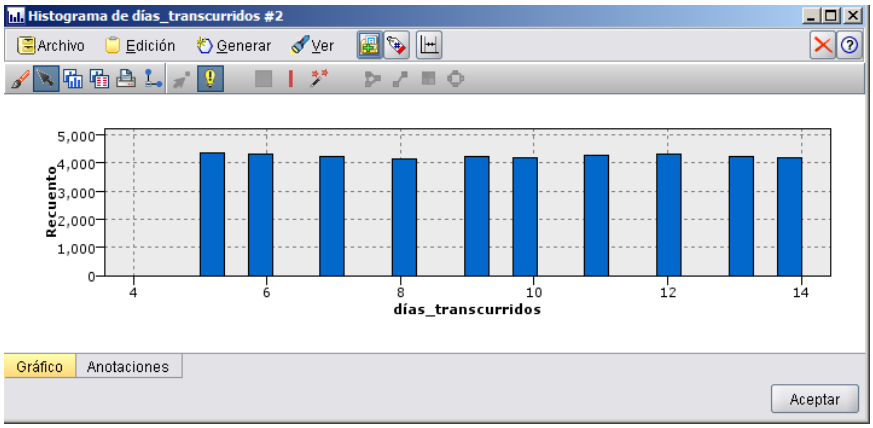


Figura III.37. Histograma de la variable días_transcurridos.

La ruta de Clementine que realiza todos estos procesos es la siguiente:

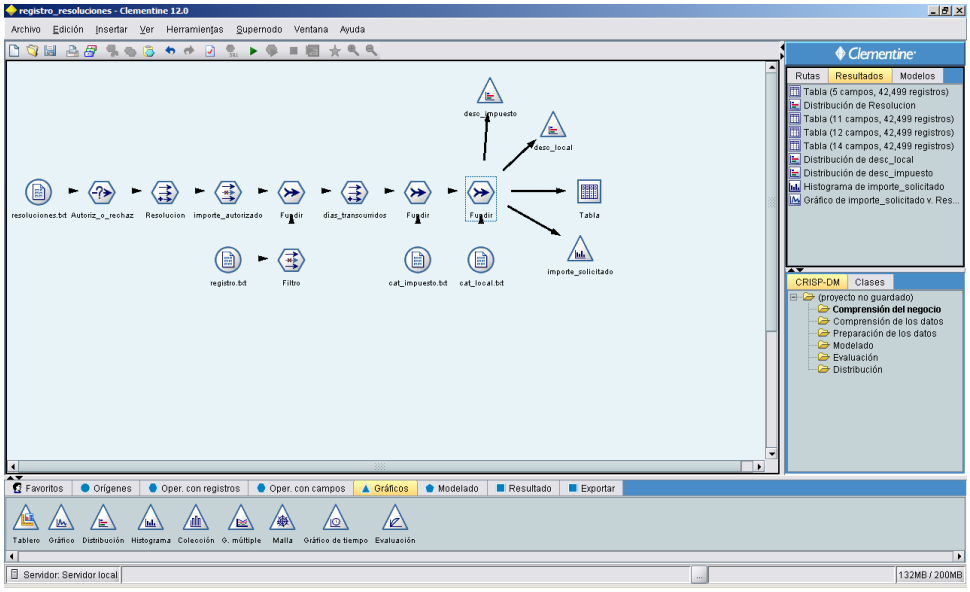


Figura III.38. Ruta en Clementine que prepara la información de devoluciones.

Con esta información se construyó una vista minable que se utilizará para la construcción del modelo predictivo.

III.2.2. Cobranza.

Para el caso de la cobranza, la tabla principal donde se registran los créditos fiscales es la tabla registro de adeudos (Registro_adeudo) esta tiene toda la historia de quienes han tenido adeudos con el fisco, sin embargo para saber el estado del crédito fiscal se deben observar las otras tablas.

En la tabla Etapas se encuentran todos los créditos registrados en las diferentes etapas por las que pasó o en las que se encuentra actualmente.

Para generar la vista minable con las bajas se deben identificar los créditos que están en la tabla bajas ya que ahí es donde está la historia de los adeudos terminados.

Se deber realizar un cruce entre la información de la tabla Bajas y la tabla Registro_adeudo y así obtener la información de registro de las bajas.

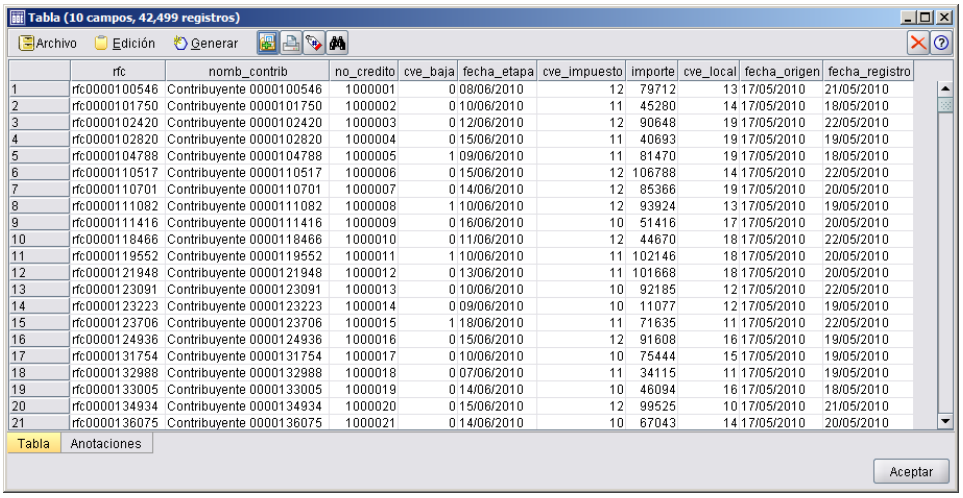


Tabla (10 campos, 42,499 registros)

Archivo Edición Generar

	rfc	nomb_contrib	no_credito	cve_baja	fecha_etapa	cve_impuesto	importe	cve_local	fecha_origen	fecha_registro
1	rfc0000100546	Contribuyente 0000100546	1000001	0	08/06/2010	12	79712	13	17/05/2010	21/05/2010
2	rfc0000101750	Contribuyente 0000101750	1000002	0	10/06/2010	11	45280	14	17/05/2010	18/05/2010
3	rfc0000102420	Contribuyente 0000102420	1000003	0	12/06/2010	12	90648	19	17/05/2010	22/05/2010
4	rfc0000102820	Contribuyente 0000102820	1000004	0	15/06/2010	11	40693	19	17/05/2010	19/05/2010
5	rfc0000104788	Contribuyente 0000104788	1000005	1	09/06/2010	11	81470	19	17/05/2010	18/05/2010
6	rfc0000110517	Contribuyente 0000110517	1000006	0	15/06/2010	12	106788	14	17/05/2010	22/05/2010
7	rfc0000110701	Contribuyente 0000110701	1000007	0	14/06/2010	12	85366	19	17/05/2010	20/05/2010
8	rfc0000111082	Contribuyente 0000111082	1000008	1	10/06/2010	12	93924	13	17/05/2010	19/05/2010
9	rfc0000111416	Contribuyente 0000111416	1000009	0	16/06/2010	10	51416	17	17/05/2010	20/05/2010
10	rfc0000118466	Contribuyente 0000118466	1000010	0	11/06/2010	12	44670	18	17/05/2010	22/05/2010
11	rfc0000119552	Contribuyente 0000119552	1000011	1	10/06/2010	11	102146	18	17/05/2010	20/05/2010
12	rfc0000121948	Contribuyente 0000121948	1000012	0	13/06/2010	11	101668	18	17/05/2010	20/05/2010
13	rfc0000123091	Contribuyente 0000123091	1000013	0	10/06/2010	10	92185	12	17/05/2010	22/05/2010
14	rfc0000123223	Contribuyente 0000123223	1000014	0	09/06/2010	10	11077	12	17/05/2010	19/05/2010
15	rfc0000123706	Contribuyente 0000123706	1000015	1	18/06/2010	11	71635	11	17/05/2010	22/05/2010
16	rfc0000124936	Contribuyente 0000124936	1000016	0	15/06/2010	12	91608	16	17/05/2010	19/05/2010
17	rfc0000131754	Contribuyente 0000131754	1000017	0	10/06/2010	10	75444	15	17/05/2010	19/05/2010
18	rfc0000132988	Contribuyente 0000132988	1000018	0	07/06/2010	11	34115	11	17/05/2010	19/05/2010
19	rfc0000133005	Contribuyente 0000133005	1000019	0	14/06/2010	10	46094	16	17/05/2010	18/05/2010
20	rfc0000134934	Contribuyente 0000134934	1000020	0	15/06/2010	12	99525	10	17/05/2010	21/05/2010
21	rfc0000136075	Contribuyente 0000136075	1000021	0	14/06/2010	10	67043	14	17/05/2010	20/05/2010

Tabla Anotaciones

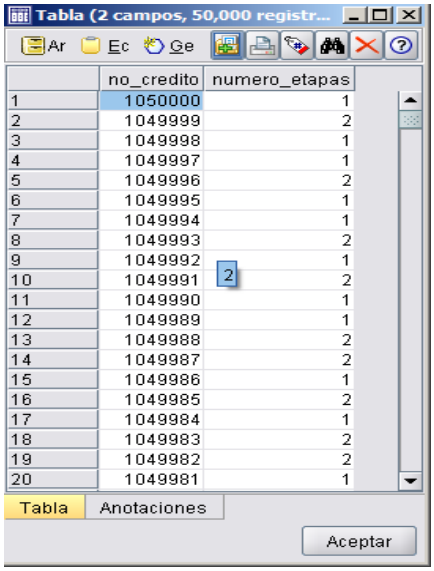
Aceptar

Figura III.39. Muestra de información del cruce entre la Bajas y Registro_adeudo.

Como se puede observar la cantidad de registros que se obtiene es la misma de la tabla bajas, esto es porque la tabla registro de adeudos tiene tanto los créditos fiscales activos como bajas, pero al cruzarlo contra la tabla de estos últimos el resultado son solo los terminados.

De la tabla de etapas sería interesante saber por cuántas distintas etapas pasa un crédito fiscal antes de ser dado de baja.

Sin embargo esta tabla tiene créditos fiscales repetidos por tantas etapas que ha pasado, por lo que primero se debe de realizar una depuración sobre los datos para no repetir la llave en la vista minable.

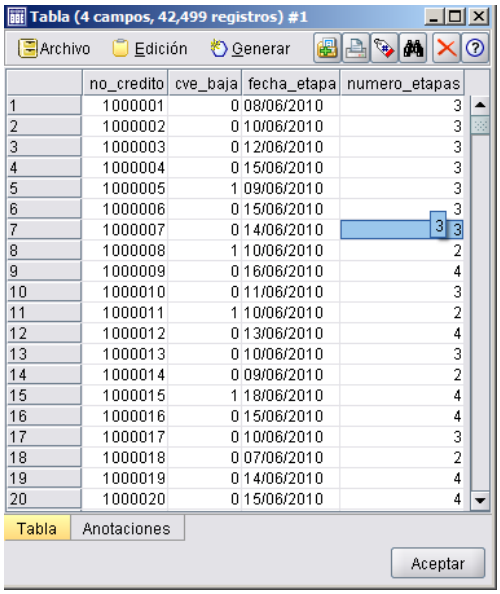


	no_credito	numero_etapas
1	1050000	1
2	1049999	2
3	1049998	1
4	1049997	1
5	1049996	2
6	1049995	1
7	1049994	1
8	1049993	2
9	1049992	1
10	1049991	2
11	1049990	1
12	1049989	1
13	1049988	2
14	1049987	2
15	1049986	1
16	1049985	2
17	1049984	1
18	1049983	2
19	1049982	2
20	1049981	1

Figura III.40. Muestra de información de créditos y número de etapas.

Con este proceso se tiene un conteo por crédito fiscal del número de etapas por las que ha pasado, sin embargo la tabla tiene tanto créditos activos como bajas, se puede observar que la tabla tiene 50 mil registros que son el total global de créditos, hay que realizar un cruce con las bajas para que se obtenga solo el conteo de los créditos que ya terminaron su ciclo.

Al realizar el cruce se tiene el siguiente resultado:

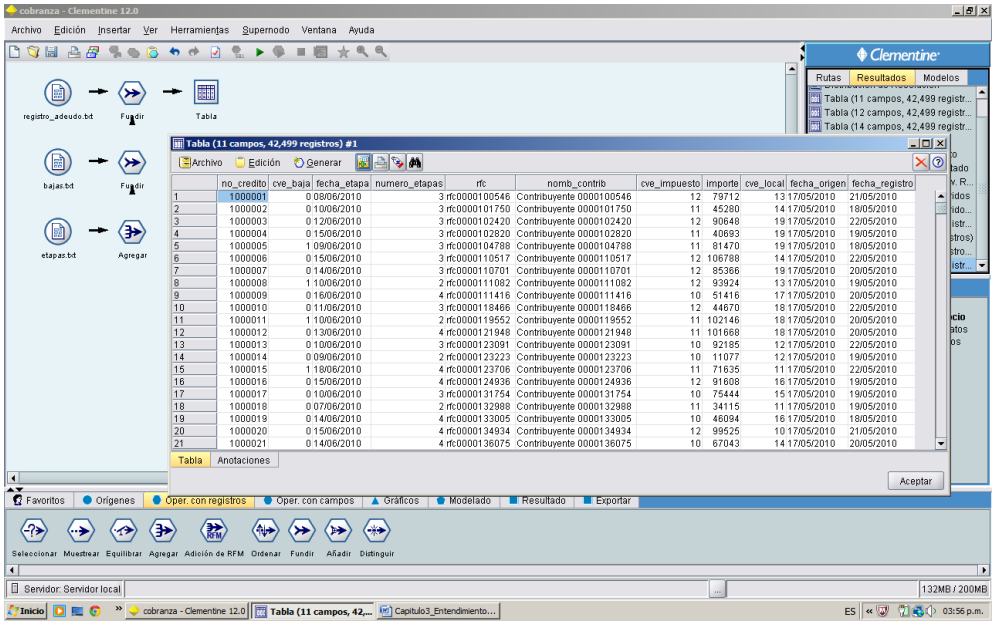


	no_credito	cve_baja	fecha_etapa	numero_etapas
1	1000001	0	08/06/2010	3
2	1000002	0	10/06/2010	3
3	1000003	0	12/06/2010	3
4	1000004	0	15/06/2010	3
5	1000005	1	09/06/2010	3
6	1000006	0	15/06/2010	3
7	1000007	0	14/06/2010	3
8	1000008	1	10/06/2010	2
9	1000009	0	16/06/2010	4
10	1000010	0	11/06/2010	3
11	1000011	1	10/06/2010	2
12	1000012	0	13/06/2010	4
13	1000013	0	10/06/2010	3
14	1000014	0	09/06/2010	2
15	1000015	1	18/06/2010	4
16	1000016	0	15/06/2010	4
17	1000017	0	10/06/2010	3
18	1000018	0	07/06/2010	2
19	1000019	0	14/06/2010	4
20	1000020	0	15/06/2010	4

Figura III.41. Muestra de información del cruce para obtener los créditos dados de baja y el número de etapas.

El número de etapas solo de los créditos dados de baja por pago o no pago.

Esta información se añadirá al cruce entre bajas y resolución adeudo:



	no_credito	cve_baja	fecha_etapa	numero_etapas	rfc	nomb_contrib	cve_impuesto	importe	cve_local	fecha_origen	fecha_registro
1	1000001	0	08/06/2010	3	rfc0000100546	Contribuyente 0000100546	12	79712	13	17/05/2010	21/05/2010
2	1000002	0	10/06/2010	3	rfc0000101750	Contribuyente 0000101750	11	45280	14	17/05/2010	18/05/2010
3	1000003	0	12/06/2010	3	rfc0000102420	Contribuyente 0000102420	12	90648	19	17/05/2010	22/05/2010
4	1000004	0	15/06/2010	3	rfc0000102820	Contribuyente 0000102820	11	40693	19	17/05/2010	19/05/2010
5	1000005	1	09/06/2010	3	rfc0000104788	Contribuyente 0000104788	11	81470	19	17/05/2010	18/05/2010
6	1000006	0	15/06/2010	3	rfc0000110517	Contribuyente 0000110517	12	106788	14	17/05/2010	22/05/2010
7	1000007	0	14/06/2010	3	rfc0000110701	Contribuyente 0000110701	12	85366	19	17/05/2010	20/05/2010
8	1000008	1	10/06/2010	2	rfc0000111082	Contribuyente 0000111082	12	93924	13	17/05/2010	19/05/2010
9	1000009	0	16/06/2010	4	rfc0000111416	Contribuyente 0000111416	10	51416	17	17/05/2010	20/05/2010
10	1000010	0	11/06/2010	3	rfc0000118466	Contribuyente 0000118466	12	44670	18	17/05/2010	22/05/2010
11	1000011	1	10/06/2010	2	rfc0000119552	Contribuyente 0000119552	11	102146	18	17/05/2010	20/05/2010
12	1000012	0	13/06/2010	4	rfc0000121948	Contribuyente 0000121948	11	101668	18	17/05/2010	20/05/2010
13	1000013	0	10/06/2010	3	rfc0000123091	Contribuyente 0000123091	10	92185	12	17/05/2010	22/05/2010
14	1000014	0	09/06/2010	2	rfc0000123223	Contribuyente 0000123223	10	11077	12	17/05/2010	19/05/2010
15	1000015	1	18/06/2010	4	rfc0000123706	Contribuyente 0000123706	11	71635	11	17/05/2010	22/05/2010
16	1000016	0	15/06/2010	4	rfc0000124936	Contribuyente 0000124936	12	91608	16	17/05/2010	19/05/2010
17	1000017	0	10/06/2010	3	rfc0000131754	Contribuyente 0000131754	10	75444	15	17/05/2010	19/05/2010
18	1000018	0	07/06/2010	2	rfc0000132988	Contribuyente 0000132988	11	34115	11	17/05/2010	19/05/2010
19	1000019	0	14/06/2010	4	rfc0000133005	Contribuyente 0000133005	10	46094	16	17/05/2010	18/05/2010
20	1000020	0	15/06/2010	4	rfc0000134934	Contribuyente 0000134934	12	99525	10	17/05/2010	21/05/2010
21	1000021	0	14/06/2010	4	rfc0000136075	Contribuyente 0000136075	10	67043	14	17/05/2010	20/05/2010

Figura III.42. Ruta con los cruces y muestra de la información creada con el cruce de las 3 fuentes.

La tabla resultante ya tiene el campo numero_etapas más los campos de las otras dos tablas principales.

Como en el caso de las devoluciones se generará la variable objetivo para la vista minable, ésta será la variable Pago, la cual tendrá el valor V (verdadero) cuando el crédito fiscal se dio de baja por pago y F (falso) cuando fue por no pago.

La configuración del nodo para la reasignación y creación de la variable es la siguiente.

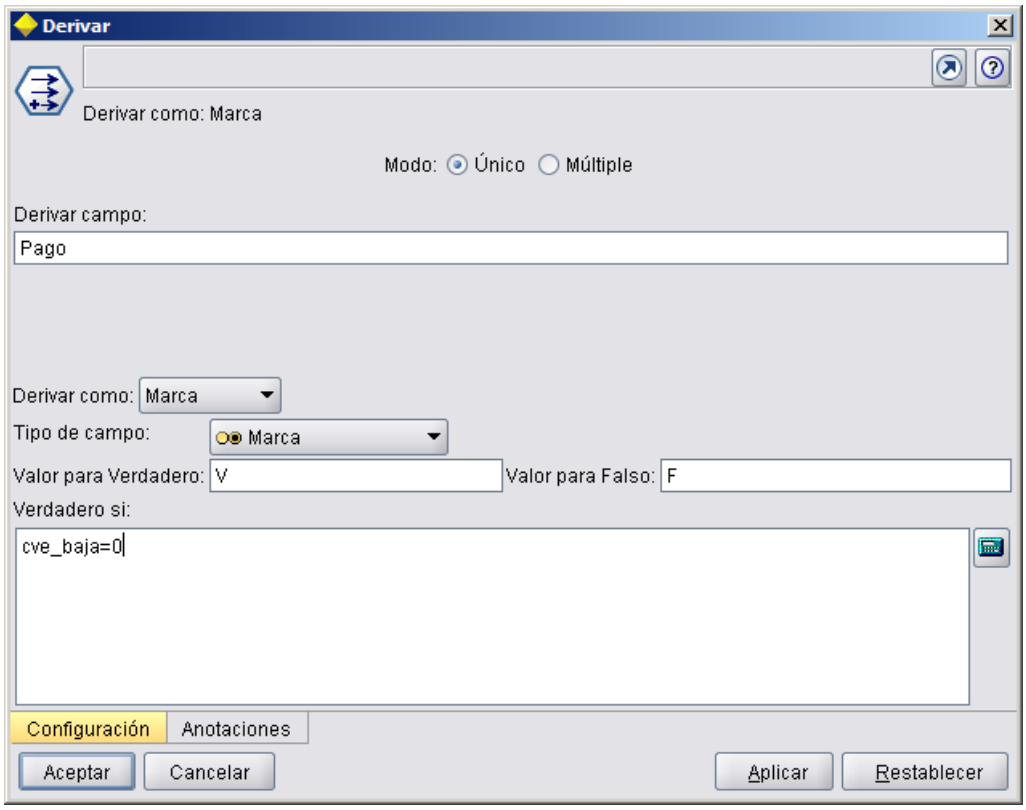


Figura III.43. Creación de la variable pago.

La vista minable ahora tendrá un campo nuevo Pago que será la variable objetivo para el modelo.

Para derivar más variables como en el caso de las devoluciones sería de interés saber el tiempo que transcurrió desde que se originó o registró el crédito hasta que se dio de baja por pago o no pago.

Falta añadir la descripción de las claves que se tiene en la vista minable como es la clave del impuesto y de local.

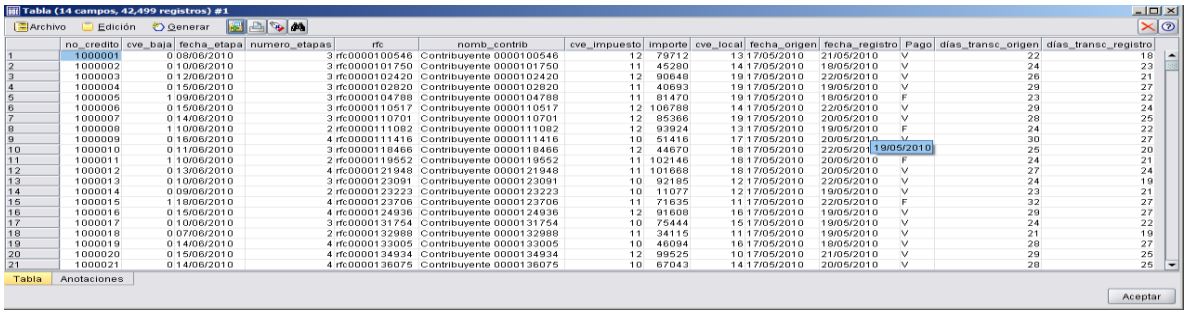


Figura III.44. Muestra de la información creada.

Se generan tres campos más a la vista minable, la variable objetivo y las diferencias de días del origen y registro del crédito.

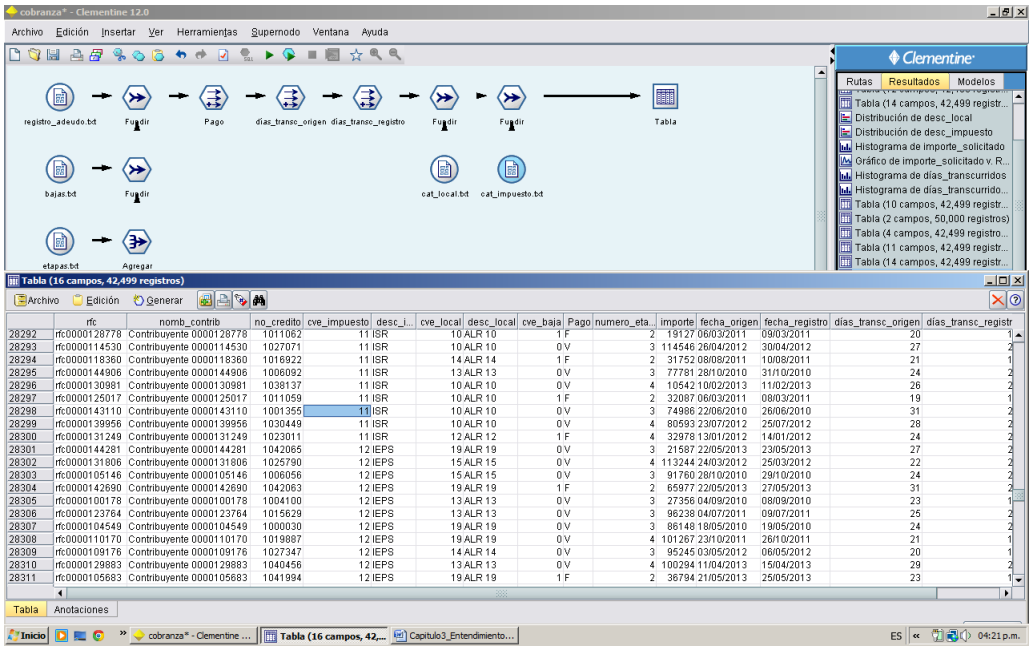


Figura III.45. Ruta con los cruces, generación de variables y muestra del resultado.

Se tiene toda la información de la vista minable, a partir de este momento se puede empezar a explorar la información contenida.

La distribución de la variable objetivo es la siguiente:

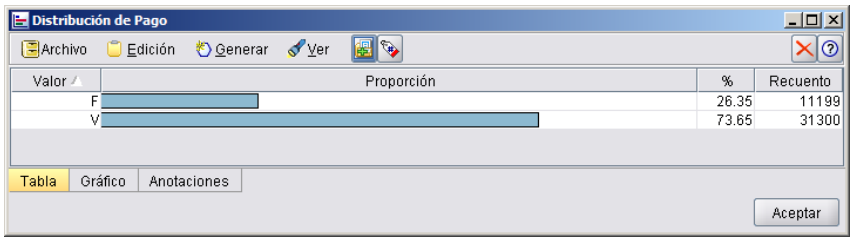


Figura III.46. Distribución de la variable objetivo Pago.

Se tiene que el pago es del 74% contra el 26% (redondeando ambas cifras) de las bajas.

La variable de impuesto muestra la siguiente distribución en sus datos:

Figura III.47. Distribución de la variable descripción de impuesto.

Se tienen 3 diferentes impuestos para los créditos fiscales los cuales tienen una distribución muy homogénea.

Para las ALR se tiene:

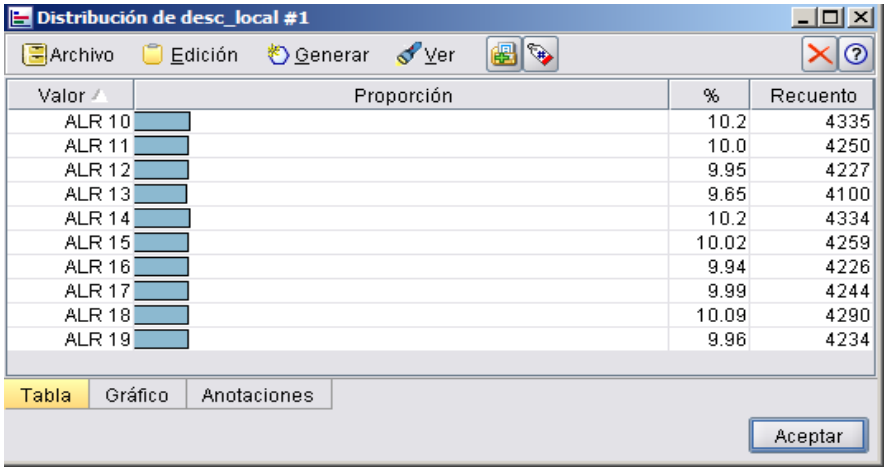


Figura III.48. Distribución de la variable descripción de local.

Se tienen 10 diferentes ALR que controlan los créditos fiscales.

El histograma del importe de los créditos fiscales es el siguiente:

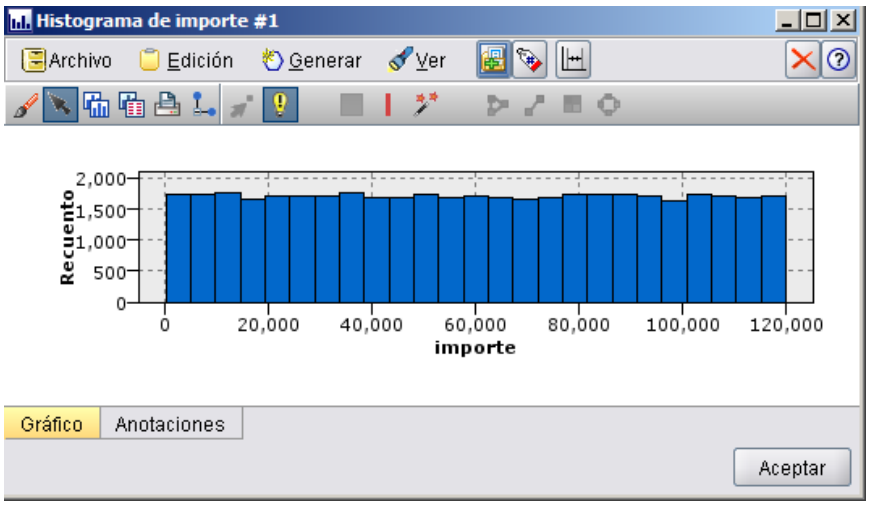


Figura III.49. Histograma de la variable importe.

El importe de los créditos fiscales va en un rango de mayor a cero y hasta 120 mil pesos.

El histograma para los días transcurridos desde que se originó el crédito es el siguiente:

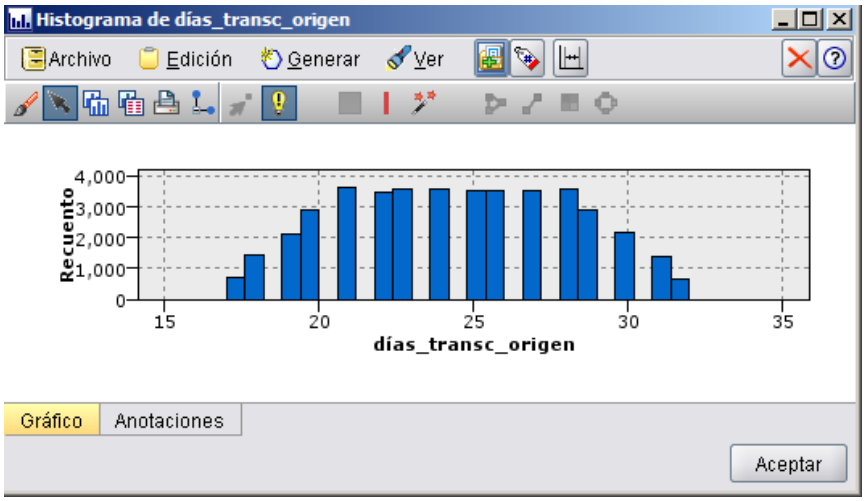


Figura III.50. Histograma de la variable días transcurridos desde que se originó el crédito.

Contiene valores de entre 17 y hasta 32 días, donde la mayoría se encuentra entre 20 y 25 días.

Para los días que transcurrieron desde el registro hasta la baja se tiene:

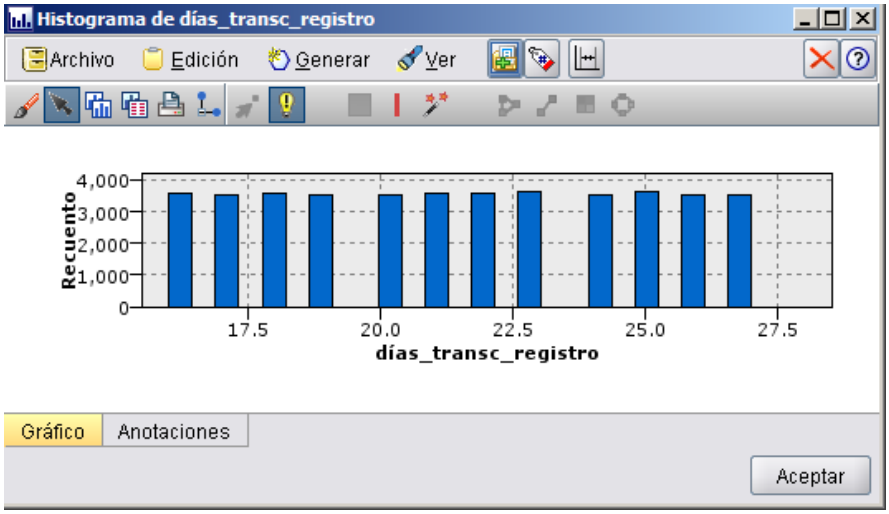


Figura III.51. Histograma de la variable días transcurridos desde el registro.

Los días transcurridos desde el registro va desde mayor a cero y hasta menor a 27 días.

La ruta de Clementine con la cual se construyó toda esta información es:

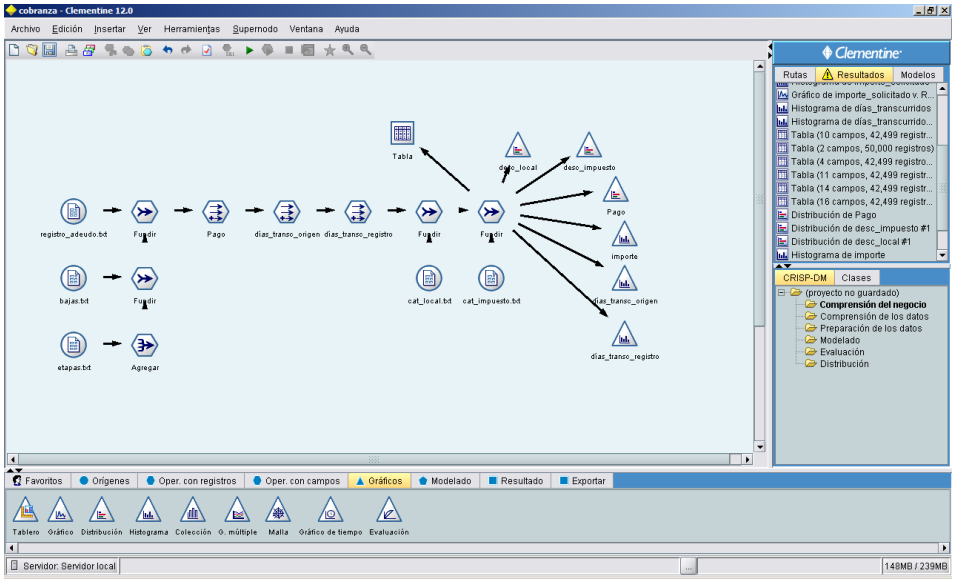


Figura III.52. Ruta de Clementine completa para la preparación de los datos para cobranza.

Resumen etapas.

En el entendimiento de los datos se revisó dónde se contenía la información y cómo estaba definida en las bases de datos, se generaron reportes de auditoría para cada tabla, se tuvo una primera aproximación al contenido de las tablas.

En la preparación de los datos se generaron las vistas minables que se utilizarán para la construcción de los modelos, se transformaron variables, se crearon nuevas y se adjuntó toda la información necesaria que se necesitará en el modelado.

El siguiente paso es la construcción del modelo con la información entendida y preparada.

III.3. Recapitulación.

En este capítulo se mostró de manera general lo que es la metodología CRISP-DM, sus etapas y su ciclo en el proceso para desarrollo de modelos.

Se revisaron los dos pasos siguientes después del entendimiento del problema.

Por un lado el entendimiento de los datos, el cual tiene la siguiente finalidad:

- Obtener la documentación de las definiciones de datos en donde se almacena la información.
- Revisar en qué plataformas se encuentra la información con la cual se trabajará.
- Revisar la estructura en la cual se encuentran los datos, por ejemplo la base de datos relacional donde se guarda la información.
- Generar reportes con la información encontrada.
- Generar los procesos necesarios para poder extraer los datos.

Por otro lado la preparación de los datos, la cual consiste de los siguientes puntos importantes:

- Obtener la información almacenada que se considere importante para el proyecto.
- Generar los procesos necesarios para juntar la información en una sola tabla.
- Formatear la información para tenerla lista para el modelado.
- Generar información adicional a partir de los datos existentes que puedan enriquecer las variables que ya existen.
- Reformatear los datos para tenerlos de una forma que puedan ser de más utilidad dentro del proceso.
- Generar una vista minable que se utilizará para el modelado del proyecto de minería de datos.

Ambos procesos son de gran importancia para el siguiente paso dentro del proyecto, del entendimiento y preparación de los datos depende la calidad de la información final, ésta redituará en mejores modelos.

Capítulo IV. Modelado y evaluación.

En el presente capítulo revisaremos lo que es el modelado, una de las partes finales de un proyecto de minería de datos.

Como ya lo hemos visto en capítulos anteriores, se entendió el problema y la información, se prepararon los datos, el siguiente paso es modelar la información con algoritmos para generar predicciones.

El modelado de los datos nos permite tener una respuesta final al problema planteado utilizando algoritmos predictivos que nos den resultados útiles, novedosos con los cuales tomar decisiones de negocio.

Como ya lo vimos en el capítulo I, existen diferentes algoritmos que se pueden utilizar en minería de datos para obtener predicciones con cierto grado de confiabilidad. Todo va a depender de la variable objetivo construida dentro de nuestra vista minable.

Algoritmos como la regresión lineal utilizan variables objetivo numéricas mientras que los árboles de decisión pueden trabajar también con tipos de datos cadena.

Se mostrará la importancia de la tipificación de las variables de entrada, la comparación de los diferentes modelos, la evaluación de éstos y al final la elección del ganador.

Dentro de este proceso es importante mencionar que en ocasiones hay que ir hacia atrás, esto porque puede ser que nos demos cuenta que nos faltan variables, o se pueden construir otras con diferentes fuentes, lo anterior con el fin de enriquecer el modelado, entonces se regresaría a la preparación de los datos.

IV.1. Devoluciones.

Se construyeron las variables de entrada y objetivo, al empezar a modelar se debe de identificar el tipo de variables con las cuales se trabajará, esto es importante porque así el algoritmo a utilizar podrá darle el tratamiento que requieren las variables.

El archivo con el que se trabajará tiene las siguientes variables:

Variables	Descripción
cve_local	Clave de la Administración Local.
cve_impuesto	Clave del impuesto.
no_solic	Número de solicitud.
importe_autorizado	Importe autorizado.
cve_estatus	Clave de estatus.
fecha_estado	Fecha de estado.
Resolucion	Resolución de la devolución.
rfc	Registro federal de contribuyentes.
nomb_contrib	Nombre del contribuyente.
importe_solicitado	Importe solicitado.
fecha_solicitud	Fecha de solicitud.
días_transcurridos	Días transcurridos.
desc_impuesto	Descripción del impuesto.
desc_local	Nombre de la administración local.

Tabla IV.1. Campos que se utilizarán para el modelado en devoluciones.

Se empieza por revisar éstos en la ruta de Clementine que se generará para este fin.

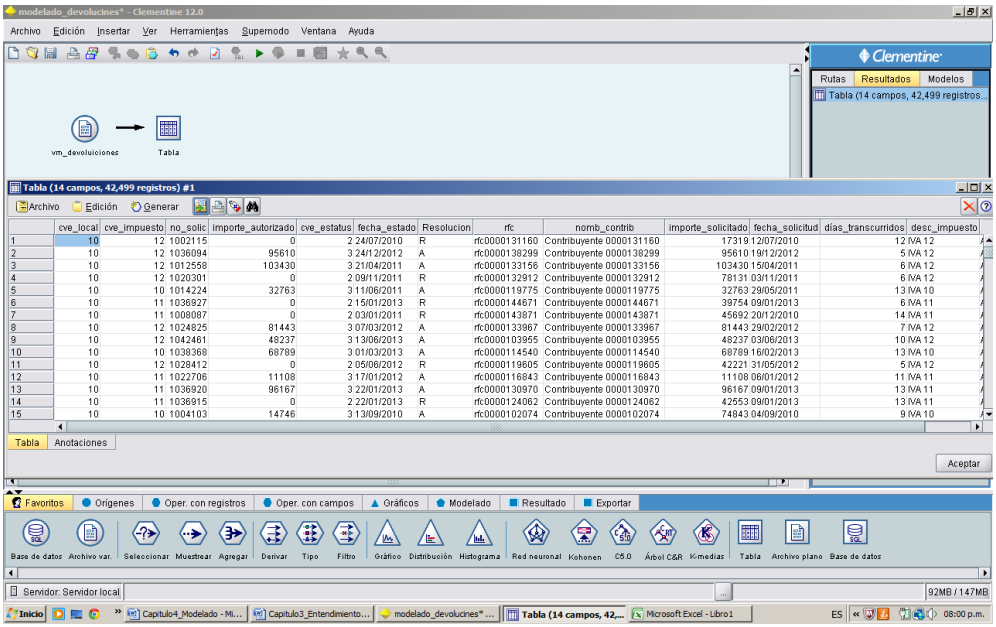


Figura IV.1. Ruta en Clementine que .genera una muestra de la información de la vista minable de Devoluciones.

Examinando el archivo se pueden ver las propiedades de las variables.

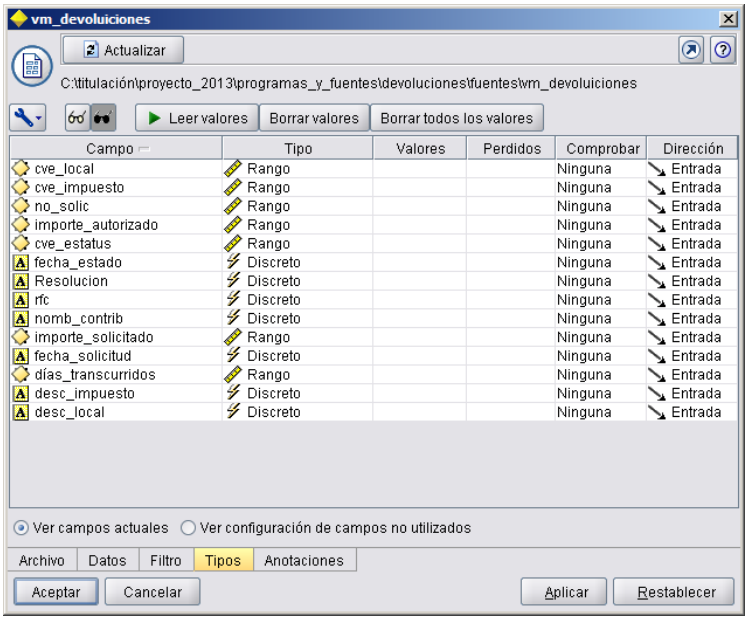


Figura IV.2. Propiedades de las variables de la vista minable de Devoluciones.

En la pestaña tipo, se puede observar cómo el minero de datos define el tipo de las variables (campo), vemos que solo se tienen Rango y discreto, esto es porque no se han leído los valores, al realizarlo se tiene el siguiente resultado.

Es importante mencionar que todo esto se está generando desde el nodo origen del archivo, se puede realizar desde aquí sin embargo no es la mejor manera de hacerlo ya que existe un nodo particular para esta tarea.

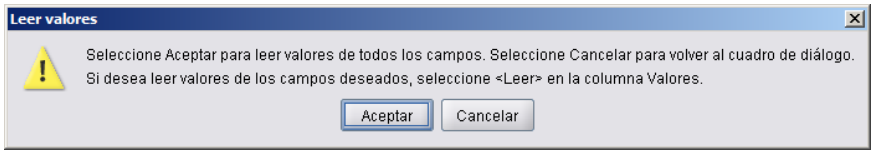


Figura IV.3. Aviso que se muestra cuando se solicita leer los valores de las variables.

La herramienta lee los valores y genera tipos para cada variable:

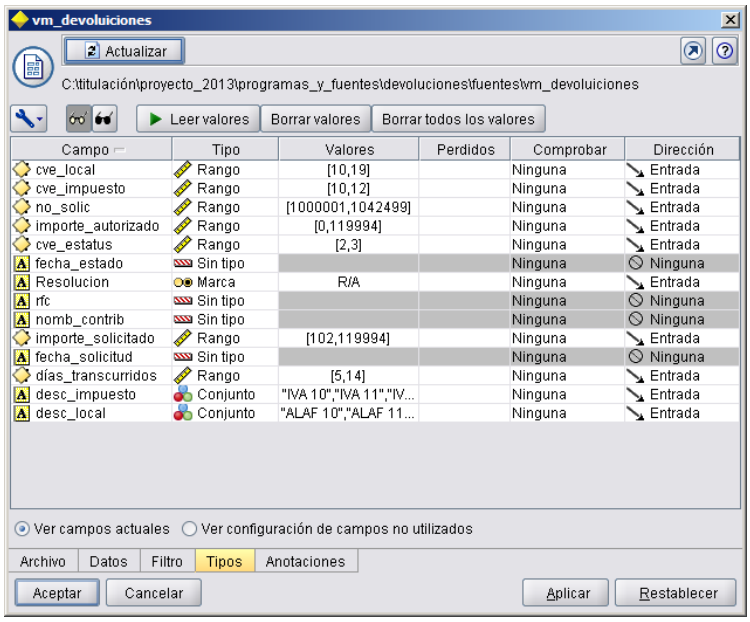


Figura IV.4. Resultado de leer los datos de la vista minable, se muestran los tipos y los valores de las variables.

Se puede observar que el tipo de dato se define a partir de los valores que contiene cada variable, no en todos los casos la clasificación es la correcta, ésta se puede cambiar manualmente.

Antes de hacer esto, es recomendable, para tener una ruta de Clementine más didáctica y mejor organizada, utilizar los nodos específicos para cada tarea, como veremos hay uno que se utiliza para leer los tipos de datos, lo que vimos anteriormente se obtiene del nodo fuente.

Sin embargo esa no es la mejor manera de hacerlo ya que el minero es una herramienta muy visual que nos permite mediante nodos secuenciales poder identificar cada paso dentro de una ruta. Para ejemplificarlo veamos cómo se ve la siguiente pantalla.

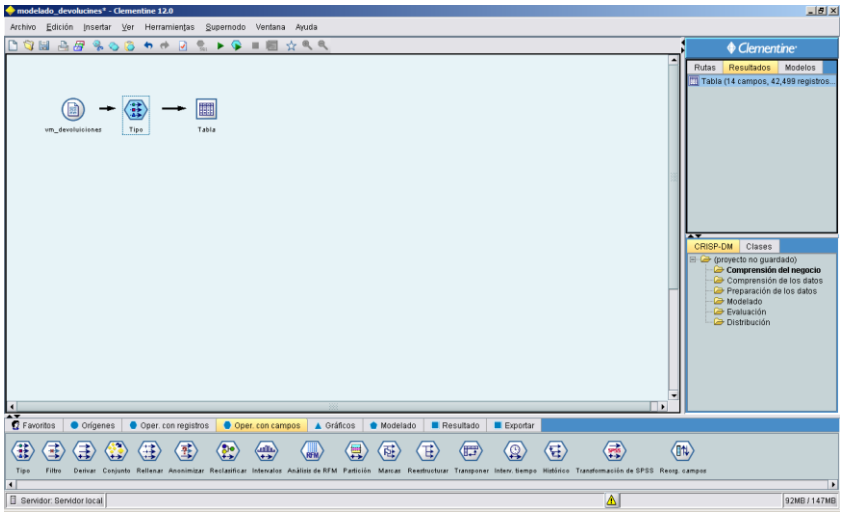


Figura IV.5. Se añade un nodo tipo al archivo, es la mejor práctica de tipificar los datos.

Al añadirle un nodo tipo al archivo fuente se observa de manera muy natural que se están tipificando los datos del archivo vm_devoluciones (vista minable devoluciones), de otra forma, si lo realizáramos en el nodo del archivo no se vería tan claro.

Estas son prácticas que deben llevarse a cabo para tener rutas que puedan ser lo más explícitas posibles y que puedan ayudar en su comprensión por ejemplo cuando una persona nueva que no esté familiarizada con el proyecto vea la ruta.

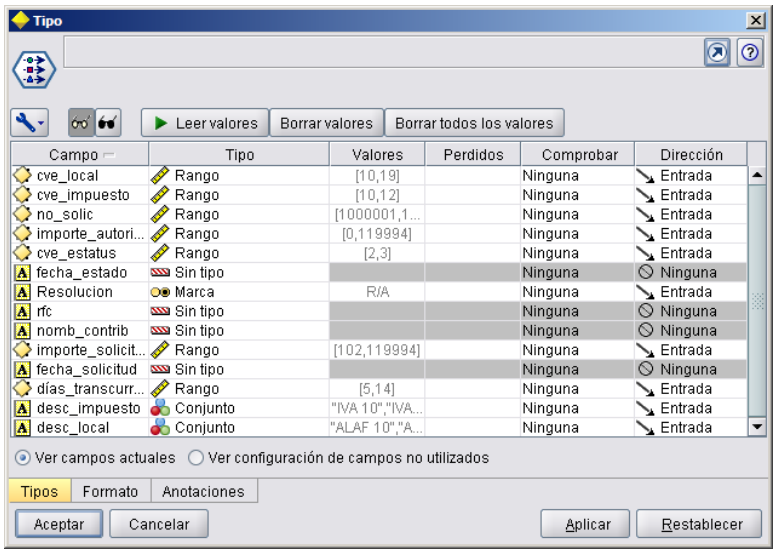


Figura IV.6. Se muestra las opciones del nodo tipo.

Vemos las características del nodo tipo son prácticamente iguales a las del nodo fuente, sin embargo el primero solo sirve para la tarea de tipificación y el segundo tiene más opciones (Archivo, Datos, Filtro).

Como ya desde el nodo fuente se leyeron los datos el nodo tipo toma la ya realizada por su antecesor. Desde este nodo corregiremos los tipos erróneos y modificaremos la dirección de las variables.

Sabemos que las variables cve_local, cve_impuesto y cve_estatus aún y cuando son números, para el modelado no deben de ser tratados como tales, es decir, son conjuntos que se utilizan para identificar administraciones locales, impuestos y estados de las devoluciones, por lo que hay que corregir de la siguiente manera:

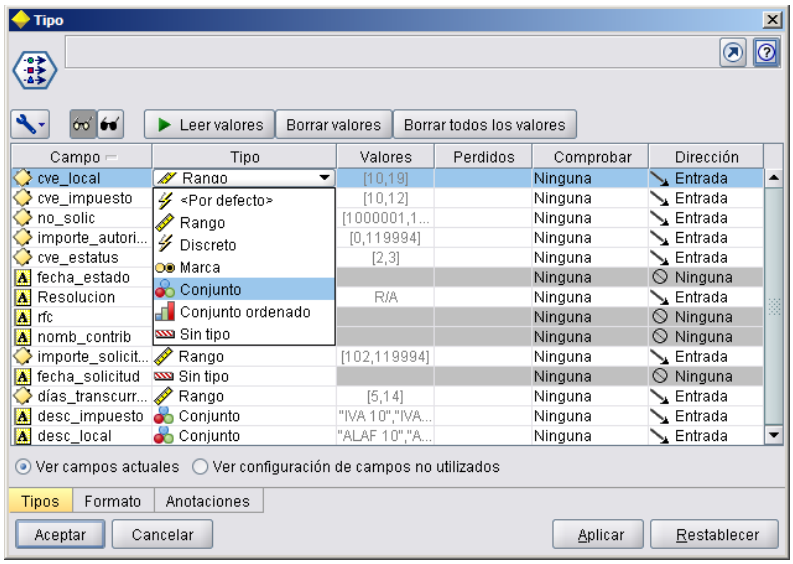


Figura IV.7. Corrección de los tipos de las variables.

Se elige la opción conjunto dentro de los valores de tipo, lo mismo para las otras dos variables mencionadas, se tiene ahora la siguiente lectura de los tipos para nuestros datos:

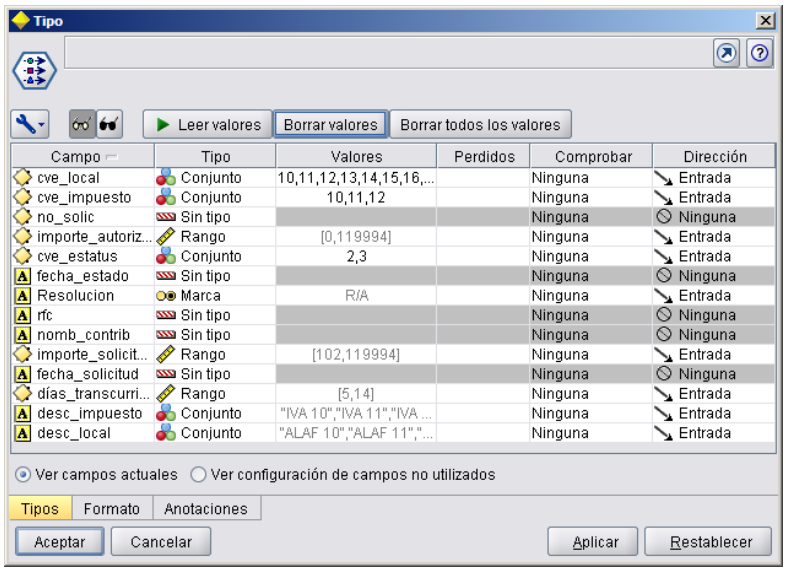


Figura IV.8. Tipificación final de las variables de la vista minable.

Podemos ver que ahora si ya se tienen los tipos correctos para todos nuestras variables.

Es importante mencionar que las fechas y los identificadores el software los clasifica “Sin tipo”, esto es correcto ya que para fines de modelado no deben incluirse variables que tengan fechas o que sean identificadores particulares de las variables, como el RFC por ejemplo.

Esto porque no tendría sentido hacer modelos particulares para identificadores, es decir imaginemos que incluimos el RFC y el modelo al final diga, si el RFC=XXFFFF entonces debe devolverse el saldo a cargo, no tendría ninguna generalidad y solo aplicaría para el caso particular.

Pongamos atención ahora en la opción “Dirección” del nodo tipo, ésta nos sirve para indicarle al nodo de modelado, las variables que necesitamos se incluyan dentro de los algoritmos, como ya lo hemos indicado tendremos variables de entrada (independientes) y objetivo (dependientes), esto se debe realizar como sigue:

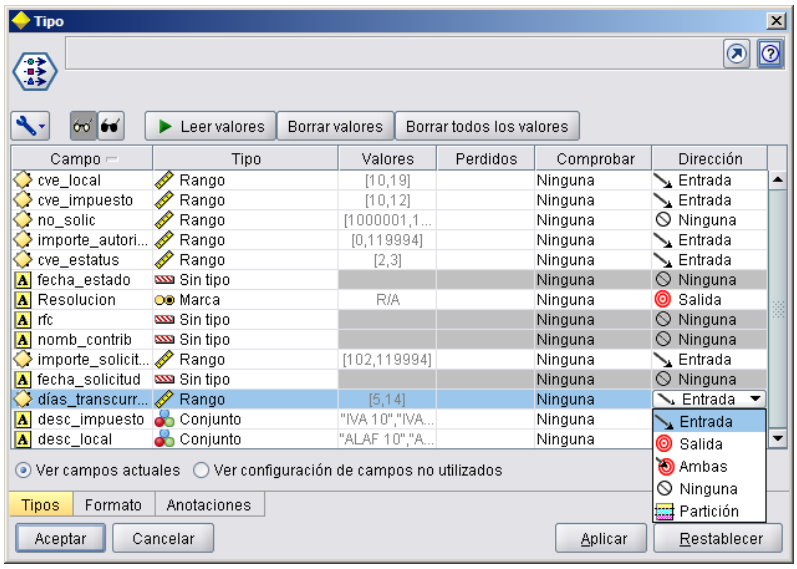


Figura IV.9. Se indica el rol que llevará a cabo cada variable de la vista minable.

Vemos que ahora ya se tienen identificadas las variables que se utilizarán en el modelado.

El siguiente paso es generar las muestras que se utilizarán, dentro del modelado deben de tenerse al menos dos muestras (en ocasiones se tienen hasta 3), una denominada de Entrenamiento y otra de Comprobación (también Validación en el caso de tres muestras).

El fin de tener estas dos muestras es la siguiente:

- **Entrenamiento.** Es la muestra con la cual se construirá el modelo, los datos se utilizarán para que el algoritmo se entrene para generar las predicciones con datos que no ha "visto", esta muestra generalmente debe ser mayor igual a la muestra de comprobación.

- **Comprobación.** Para evaluar la efectividad de nuestro modelo sin todavía utilizar los datos reales con los que al final se ejecutará, necesitamos saber qué tan bueno es el desempeño del modelo con datos nuevos, la ventaja de esta muestra es que ya sabemos lo que ocurrió al final, por lo que podremos evaluar el desempeño y será un indicador muy importante para decidir cuál es el modelo ganador.

Para realizar esta tarea se tiene un nodo particular de Clementine, el nodo Partición.

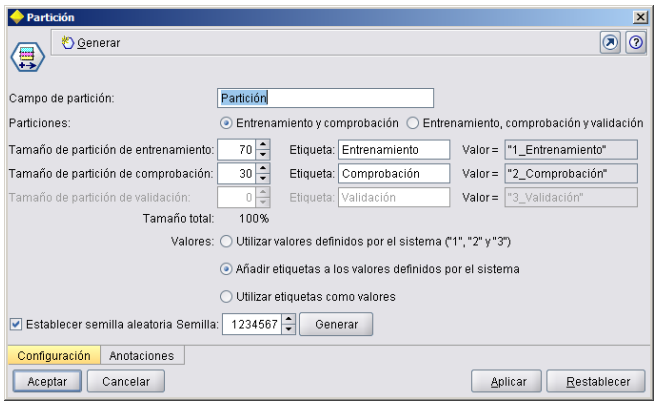


Figura IV.10. Configuración del nodo partición para generar la muestra de entrenamiento y comprobación.

Establecemos una partición de 70% para el entrenamiento y el 30% para la comprobación, el nodo se insertará después del nodo origen y antes que el nodo tipo, esta es una buena práctica dejar siempre todos los nodos antes del Tipo para que solo se tipifiquen los datos una sola vez,

Nuestra ruta queda como sigue:

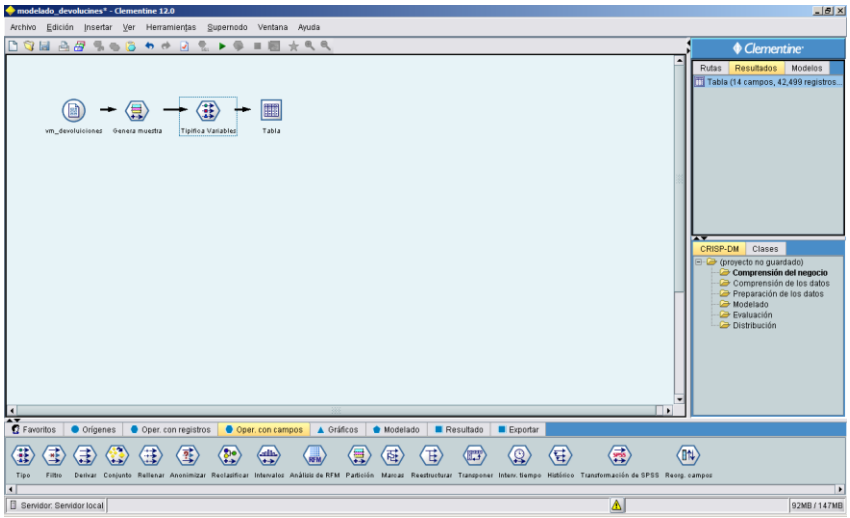


Figura IV.11. Ruta en Clementine con los pasos realizados.

Se puede observar que los nodos partición y tipo ya tienen una descripción más personalizada dentro de la ruta (“Genera muestra” y “Tipifica Variables”), esto es una buena práctica porque así la ruta puede ser más fácilmente leída.

El nodo tipo ahora tiene una nueva variable la cual tiene un tipo Marca (solo dos valores) y una Dirección -> Partición.

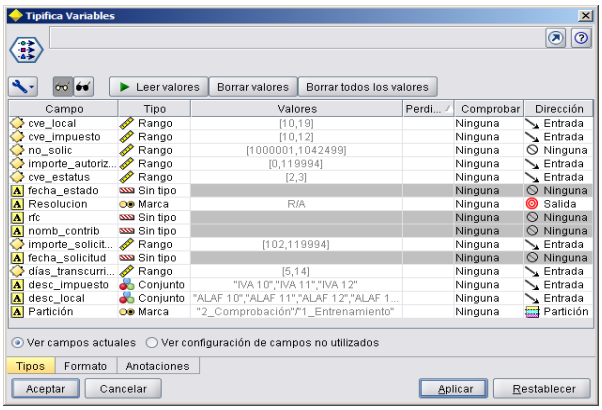


Figura IV.11. Muestra la inclusión de la variable Partición con su Tipo y Dirección.

Hasta este momento aún no hemos generado modelos, ya teniendo bien tipificados los datos y generado las muestras, se puede empezar con el modelado.

Existen diferentes algoritmos que utilizan una variable objetivo (salida) de tipo cadena, en nuestro caso Resolucion con valores R (Rechazado) o A (autorizado).

Clementine tiene un nodo que ayuda mucho para tener una primera aproximación a los modelos que se pueden utilizar, el nodo se llama: Clasificador binario, éste evalúa automáticamente todos los modelos que aplican para una variable objetivo con solo dos valores (binario).

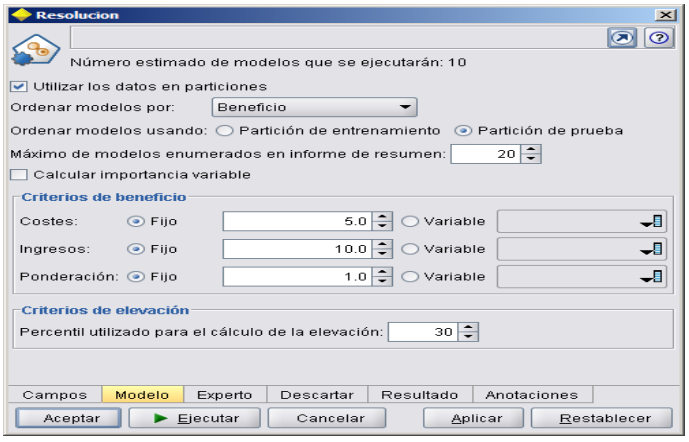


Figura IV.12. Configuración del nodo Clasificador binario.

Se muestra las opciones por default del nodo, al ejecutarlo se tienen los siguientes resultados:

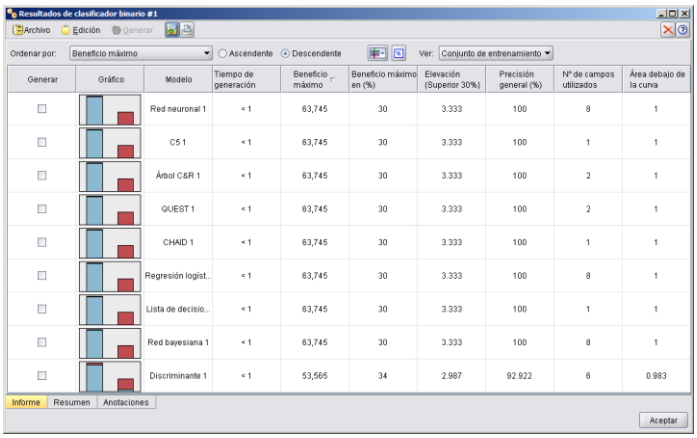


Figura IV.13. Resultados de los modelos generados por el Clasificador binario.

El nodo genera los posibles modelos a utilizarse compitiendo entre cada uno y generando los resultados de los ganadores por diferentes medidas de evaluación de cada uno, por ejemplo para este caso se tienen ordenados por el “Beneficio máximo”.

La descripción de los criterios de los métodos de evaluación son los siguientes (Fuente: ayuda de SPSS Clementine 12):

- **Precisión global.** Porcentaje de registros pronosticados correctamente por el modelo respecto al número total de registros.
- **Área debajo de la curva.** Proporciona el índice de rendimiento de un modelo. Cuanto más se encuentre la curva sobre la línea de referencia, más exacta será la prueba.
- **Beneficio (acumulado).** Suma de los beneficios de los percentiles acumulados (clasificados en términos de confianza para el pronóstico), calculados en base a los costes, ingresos y criterios de ponderación especificados.
- **Elevación (acumulado).** Tasa de aciertos en cantidades acumuladas con respecto a la muestra global (donde los cuantiles se clasifican en función de la confianza para el pronóstico).

Al observar los datos de evaluación de los modelos, encontramos que la precisión es exacta para la mayoría, esto nos da una pista que muy posiblemente tenemos variables que están completamente correlacionadas con la variable objetivo.

Para revisar esta situación generaremos un análisis de correlaciones con el nodo Estadísticos, éste solo puede ocuparse para las variables numéricas, el resultado es el siguiente:

Correlaciones de Pearson			
cve_local	cve_impuesto	0.001	Débil
	no_solic	-0.001	Débil
	importe_autorizado	0.005	Débil
	cve_estatus	0.004	Débil
	importe_solicitado	0.005	Débil
	días_transcurridos	0.005	Débil
cve_impuesto	cve_local	0.001	Débil
	no_solic	-0.003	Débil
	importe_autorizado	0.003	Débil
	cve_estatus	0.002	Débil
	importe_solicitado	0.008	Débil
	días_transcurridos	0.004	Débil
no_solic	cve_local	-0.001	Débil
	cve_impuesto	-0.003	Débil
	importe_autorizado	-0.004	Débil
	cve_estatus	-0.005	Débil
	importe_solicitado	0.001	Débil
	días_transcurridos	-0.002	Débil
importe_autorizado	cve_local	0.005	Débil
	cve_impuesto	0.003	Débil
	no_solic	-0.004	Débil
	cve_estatus	0.663	Fuerte
	importe_solicitado	0.588	Fuerte
	días_transcurridos	0.005	Débil
cve_estatus	cve_local	0.004	Débil
	cve_impuesto	0.002	Débil
	no_solic	-0.005	Débil
	importe_autorizado	0.663	Fuerte
	importe_solicitado	-0.002	Débil
	días_transcurridos	0.001	Débil
importe_solicitado	cve_local	0.005	Débil
	cve_impuesto	0.008	Débil
	no_solic	0.001	Débil
	importe_autorizado	0.588	Fuerte
	cve_estatus	-0.002	Débil
	días_transcurridos	0.011	Fuerte
días_transcurridos	cve_local	0.005	Débil
	cve_impuesto	0.004	Débil
	no_solic	-0.002	Débil
	importe_autorizado	0.005	Débil
	cve_estatus	0.001	Débil
	importe_solicitado	0.011	Fuerte

Figura IV.14. Resultado del análisis de correlaciones de las variables de entrada para el modelo.

Las variables fuertemente correlacionadas son:

Días transcurridos con importe solicitado. Tiene una correlación fuerte positiva, lo que indica que entre más grande son los días transcurridos más grande es el importe solicitado y viceversa, esto es intuitivamente normal, ya que los importes grandes deben de tener una mayor atención por las áreas fiscalizadoras.

Importe solicitado con importe autorizado. También se puede explicar porque en los casos que se autoriza una devolución en muchos casos el importe es igual al solicitado, solo en los casos en los que se rechaza entonces el autorizado es igual a cero.

Cve de estatus con importe autorizado. Tiene la correlación más fuerte de todas las variables, si recordamos en el capítulo anterior al entender los datos encontramos que la variable objetivo Resolución se genera a partir de la clave de estatus, por lo que es una variable que está completamente correlacionada, así mismo el importe autorizado, cuando es cero quiere decir que no se autorizó la devolución y cuando es mayor a cero entonces se autorizó.

Estas últimas variables son un ejemplo muy representativo del cuidado que debe tenerse al elegir las variables de entrada, deben elegirse de modo que tengan una independencia real contra la respuesta que se está buscando.

El importe solicitado y los días transcurridos deberán ser excluyentes al generar los modelos, es decir cuando se incluya una la otra no deberá entrar al modelo y viceversa.

Al ejecutar nuevamente el clasificador binario con la variable importe solicitado se obtienen los siguientes resultados:

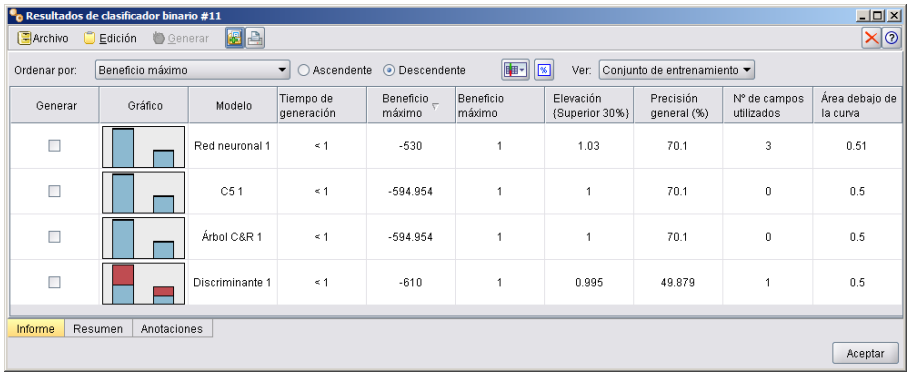


Figura IV.15. Resultado del Clasificador binario con la variable Importe solicitado.

Comparando ahora utilizando la variable días transcurridos se tiene:

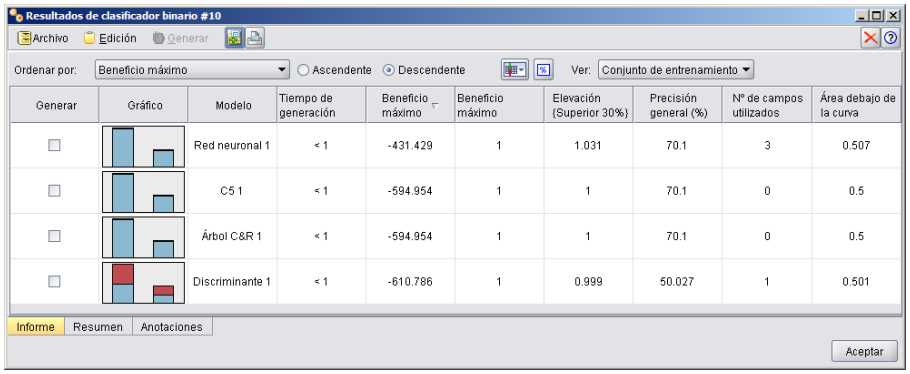


Figura IV.16. Resultado del Clasificador binario con la variable días transcurridos.

Realmente no se observa una notoria mejora incluyendo alguna, otro método para tratar de encontrar las variables más significativas es la regresión, para este caso en que la variable objetivo es cadena se puede utilizar la logística con únicamente estas dos variables y así encontrar la más significativa.

Se utilizará el nodo Logística para realizar esta tarea, la configuración es la siguiente:

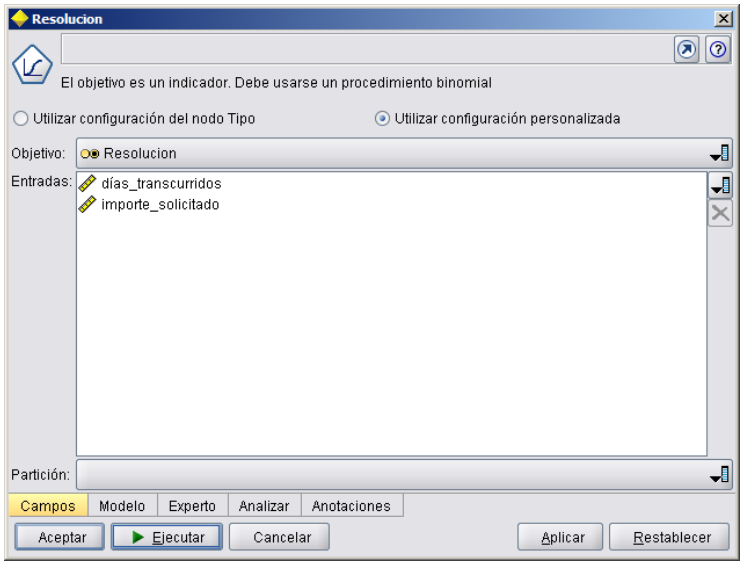


Figura IV.16. Configuración del nodo Regresión Logística para encontrar las variables más significativas.

Los resultados para la variable objetivo Resolución con el importe solicitado y los días transcurridos es el siguiente:

Resumen del procesamiento de los casos

		N	Porcentaje marginal
Resolucion	A	20793	70.10%
	R	8869	29.90%
Válidos		29662	100.00%
Perdidos		0	
Total		29662	
Subpoblación		29303(a)	

a. La variable dependiente sólo tiene un valor observado en 29168 (99.5%) subpoblaciones.

Información del ajuste del modelo

Modelo	Criterio de ajuste del modelo	Contrastes de la razón de verosimilitud		
	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersección	36000.652			
Final	36000.624	0.028	2	0.986

Pseudo R-cuadrado

Cox y Snell	0
Nagelkerke	0
McFadden	0

Estimaciones de los parámetros

Resolucion(a)		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
								Límite inferior	Límite superior
R	Intersección	-0.857	0.049	308.74	1	0			
	días_transcurridos	0.001	0.004	0.025	1	0.874	1.001	0.992	1.009
	importe_solicitado	0	0	0.003	1	0.955	1	1	1

a. La categoría de referencia es: A.

Figura IV.17. Resultados de la regresión logística.

Podemos observar que la significancia es ligeramente mayor para el importe solicitado por lo que será finalmente ésta la que se utilizará para el modelado.

Nuestro nodo tipo finalmente quedará de la siguiente forma:

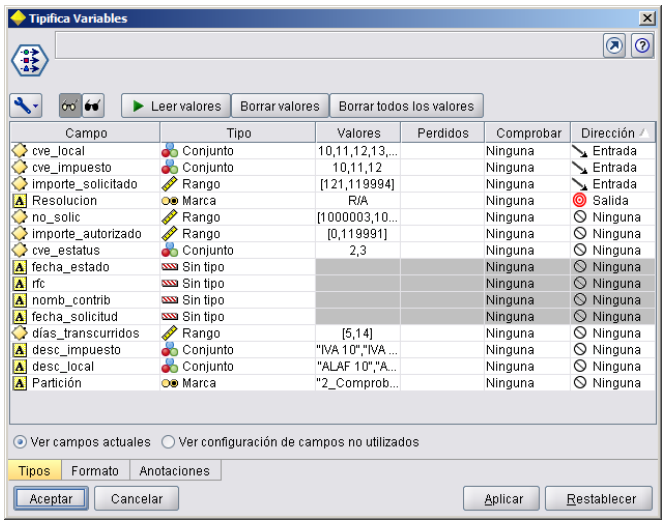


Figura IV.18. Nodo tipo con las variables finales para generar el modelo.

Con tres variables de entrada y una de salida.

Al ejecutar nuevamente el clasificador binario para estas variables se tiene:

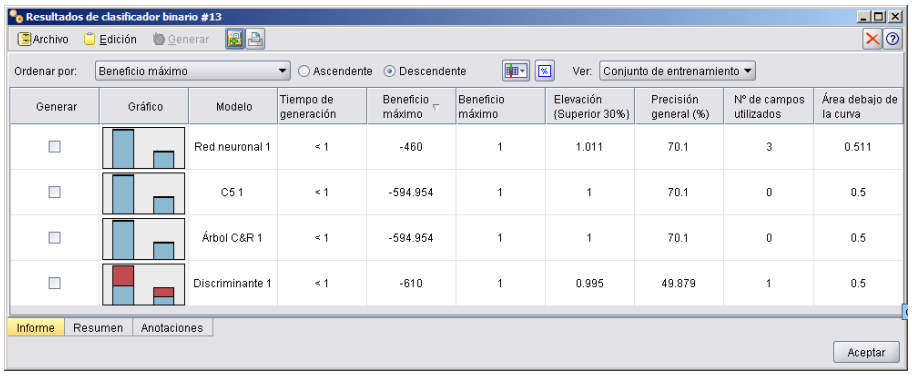


Figura IV.19. Resultado del Clasificador binario incluyendo las variables finales.

Vemos que la precisión de los modelos ya no es exacta, esto es un comportamiento normal ya que los modelos que se ajustan perfectamente generalmente tienen errores conceptuales.

Al generar los modelos de árbol de decisión C5.1 y C&R 1 para visualizar los diagramas generados (la red neuronal y discriminante generan diagramas para visualizar intuitivamente los modelos).

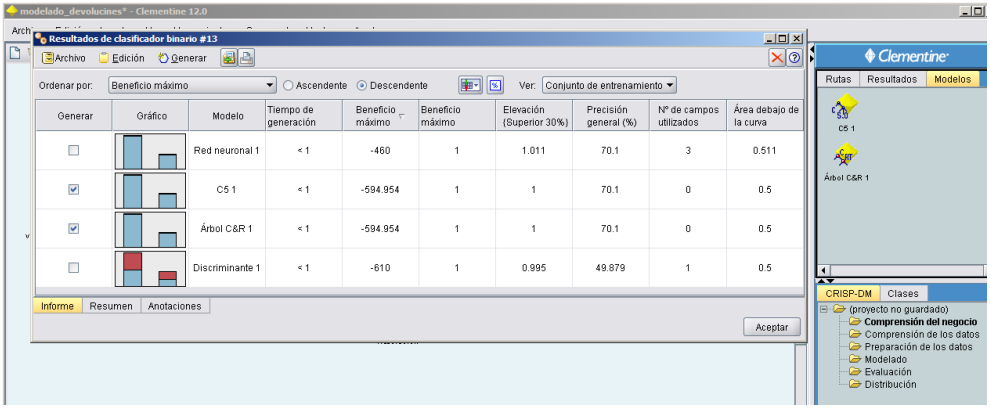


Figura IV.20. Generación de los modelos de árbol C5.1 y C&R a la paleta de modelos en Clementine.

Podemos observar lo siguiente:

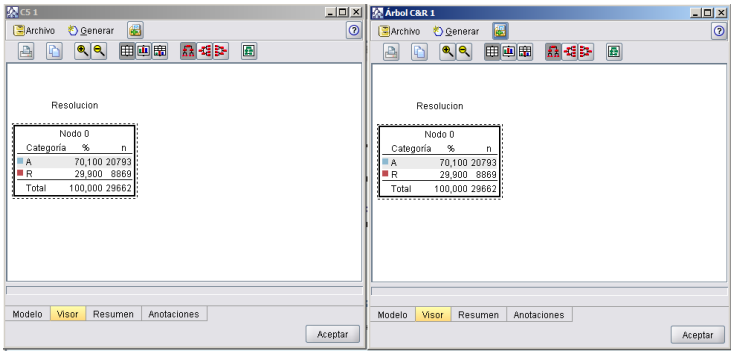


Figura IV.21. Visualización de los modelos de árbol de decisión.

No generan ningún árbol de decisión, es decir, no se necesita un modelo para predecir lo que pasará con las variables independientes, esto puede deberse a que la muestra está completamente cargada hacia A (autorizado).

Se autoriza el 70% y se rechaza el 30% aproximadamente, la muestra de la información está desbalanceada, esto se puede remediar utilizando algoritmos que balanceen la muestra, existe en Clementine un nodo para realizar esto:

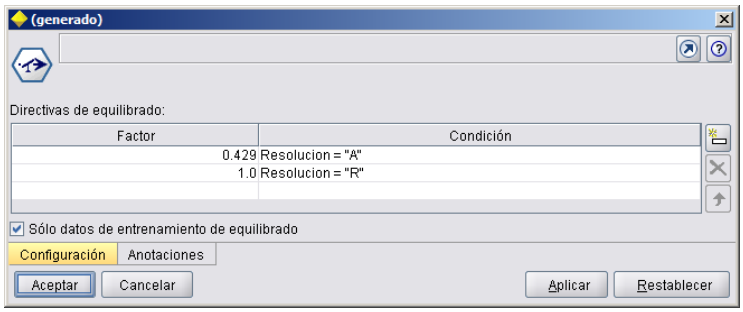


Figura IV.22. Configuración del nodo para balancear la variable objetivo.

El balanceo de la muestra puede ser en aumento, es decir se generan valores ficticios de los casos con menos registros o de disminución cuando se quitan de manera aleatoria los registros que más se tienen, es recomendable hacer esto último para no generar datos que en la realidad no existen.

En la figura de arriba (Figura IV.22.) se muestra cómo quedaría el nodo balancear muestra.

Ejecutando nuevamente el clasificador binario tenemos:

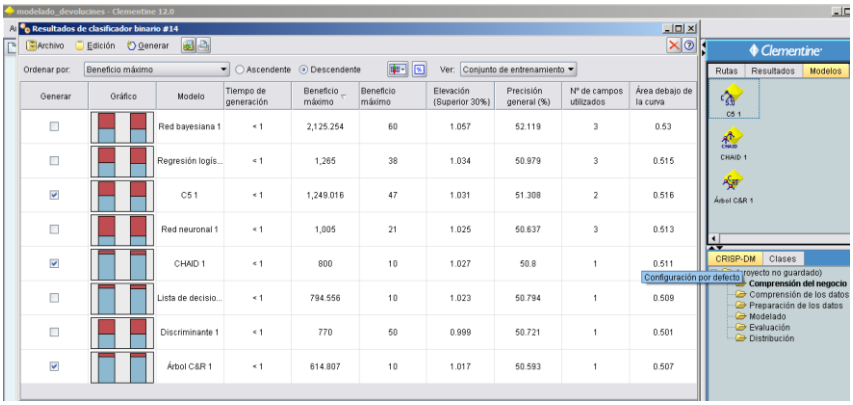


Figura IV.23. Resultados del Clasificador binario para la variable objetivo balanceada y generación de los modelos de árbol.

Revisando nuevamente solo los árboles de decisión vemos para el C5.

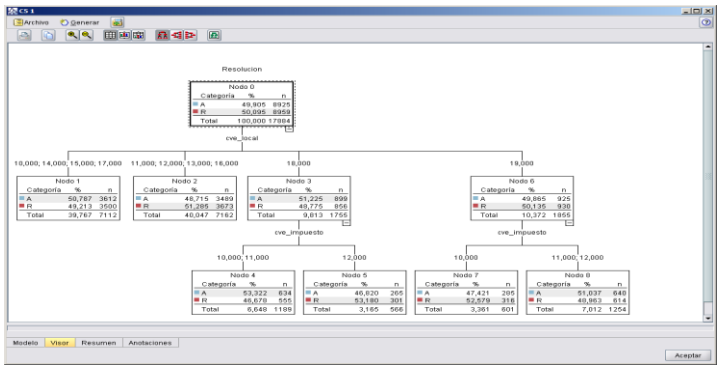


Figura IV.24. Visor del resultado para el árbol de decisión C5.

Para Chaid.

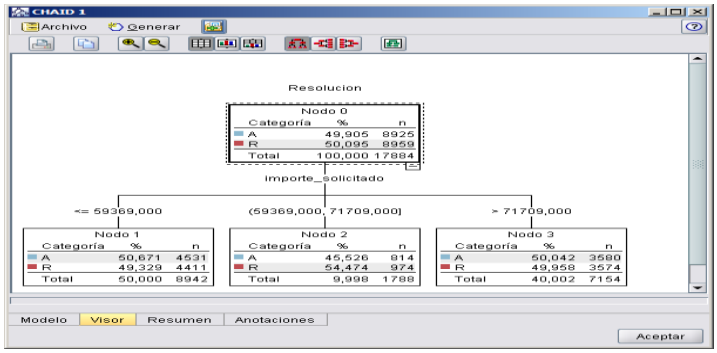


Figura IV.25. Visor del resultado para el árbol de decisión CHAID.

Y para C&R.

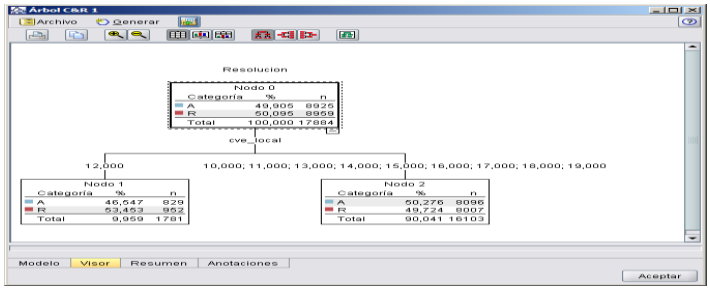


Figura IV.26. Visor del resultado para el árbol de decisión C&R.

Como se puede observar los algoritmos ya encuentran relaciones entre las variables para generar un resultado, esto nos indica que el balanceo de la muestra fue una decisión acertada porque ya podemos empezar a observar resultados interesantes como:

Para diferentes algoritmos de árboles de decisión las variables que utilizan para generarlos son diferentes en cada uno, para C&R es la clave de la local, para Chaid es el importe solicitado y para C5 son tanto la clave de la local como la clave de impuesto.

El clasificador binario encuentra 7 posibles modelos que pueden ser utilizados, sin embargo el análisis de discriminante se omitirá ya que éste solo utiliza por definición variables independientes del tipo rango, nosotros tenemos solo 3 variables para predecir el valor de resolución y dos son de tipo conjunto, así que estaríamos descartando más de la mitad de las variables de entrada.

Para encontrar cuál es el modelo que mejor se comporta para nuestro problema debemos iniciar evaluando la muestra de comprobación, recordemos que estos datos no los ha “visto” el modelo,

pero si sabemos cuál es el resultado final de esta muestra, entonces podremos evaluar contra el pasado.

Se utilizará el nodo Análisis de Clementine, la configuración para cada uno de los modelos es la siguiente:

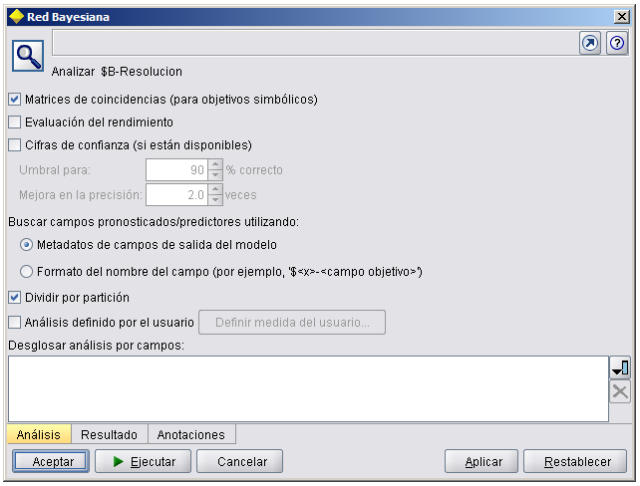


Figura IV.27. Configuración del nodo Análisis para la Red Bayesiana.

Como ejemplo se muestra el resultado del nodo Análisis para el modelo Red bayesiana.

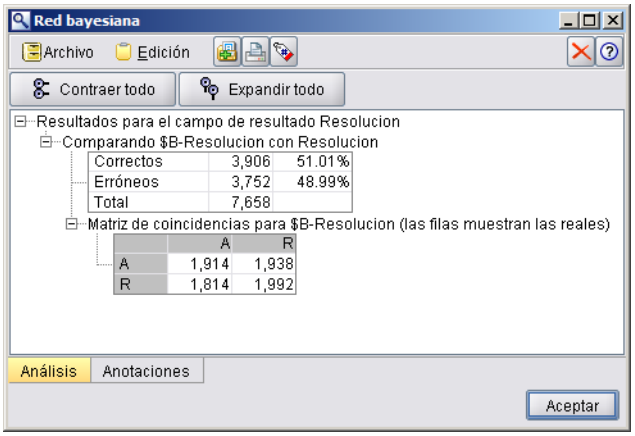


Figura IV.28. Resultados del nodo Análisis para la Red Bayesiana.

Aplicando la misma configuración para cada modelo se muestra la siguiente tabla resumen comparando la variable objetivo contra lo observado (se omite la matriz de coincidencias ya que ésta es la suma de las diagonales):

Resultados de los modelos, muestra de comprobación												
Clasificación	C&R		C5		L Decisiones		Chaid		Red bayesiana		R Logística	
	Casos	%	Casos	%	Casos	%	Casos	%	Casos	%	Casos	%
Correctos	3,879	50.65%	3,876	50.61%	400	5.22%	3,889	50.78%	3,906	51.01%	3,857	50.37%
Erróneos	3,779	49.35%	3,782	49.39%	7,258	94.78%	3,769	49.22%	3,752	48.99%	3,801	49.63%
Total Casos	7,658											

Figura IV.29. Resumen del análisis de los resultados de cada modelo.

La Red bayesiana es el modelo que mejor desempeño tiene de los 6 comparados, Lista de decisiones es el modelo con el desempeño más bajo.

El modelo ganador para nuestro problema es la Red bayesiana. Los resultados del algoritmo son los siguientes.

Para la gráfica se tiene:

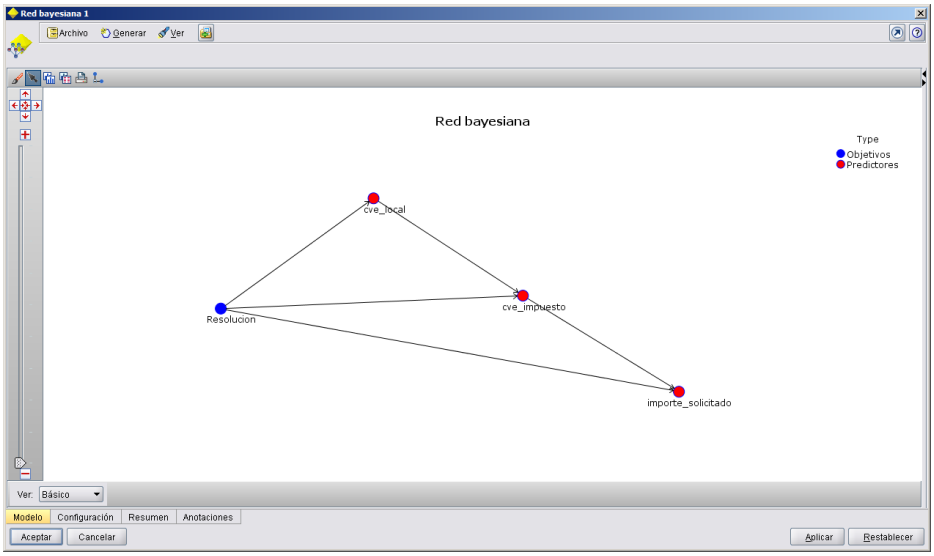


Figura IV.29. Gráfica del modelo Red bayesiana.

Las probabilidades condicionales para cada variable se muestra a continuación:

- Clave de local.

Probabilidades condicionales de cve_local

Parentales	Probabilidad									
Resolucion	10	11	12	13	14	15	16	17	18	19
R	0.097	0.104	0.106	0.097	0.097	0.097	0.102	0.098	0.095	0.103
A	0.098	0.104	0.092	0.096	0.098	0.106	0.097	0.101	0.1	0.103

Figura IV.30. Probabilidades condicionales del modelo Red bayesiana para la variable Clave de local.

- Clave de impuesto.

Probabilidades condicionales de cve_impuesto

Parentales		Probabilidad		
cve_local	Resolucion	10	11	12
10	R	0.339	0.32	0.339
10	A	0.344	0.315	0.34
11	R	0.319	0.363	0.317
11	A	0.344	0.332	0.323
12	R	0.332	0.343	0.323
12	A	0.341	0.314	0.343
13	R	0.316	0.361	0.321
13	A	0.332	0.353	0.313
14	R	0.353	0.323	0.323
14	A	0.335	0.319	0.345
15	R	0.34	0.316	0.342
15	A	0.327	0.33	0.341
16	R	0.334	0.328	0.336
16	A	0.337	0.311	0.351
17	R	0.337	0.333	0.328
17	A	0.313	0.365	0.32
18	R	0.323	0.324	0.351
18	A	0.354	0.35	0.294
19	R	0.339	0.333	0.326
19	A	0.308	0.348	0.343

Figura IV.31. Probabilidades condicionales del modelo Red bayesiana para la variable Clave de impuesto.

- Importe solicitado.

Probabilidades condicionales de importe_solicitado

Parentales		Probabilidad				
cve_impuesto	Resolucion	<= 24,095.6	24,095.6 ~ 48,070.2	48,070.2 ~ 72,044.8	72,044.8 ~ 96,019.4	> 96,019.4
10	R	0.201	0.21	0.203	0.205	0.179
10	A	0.205	0.214	0.178	0.204	0.197
11	R	0.208	0.19	0.199	0.206	0.195
11	A	0.203	0.204	0.193	0.197	0.2
12	R	0.197	0.194	0.203	0.192	0.211
12	A	0.189	0.215	0.201	0.195	0.198

Figura IV.32. Probabilidades condicionales del modelo Red bayesiana para la variable Importe solicitado.

La importancia de las variables para la Red bayesiana son los siguientes:

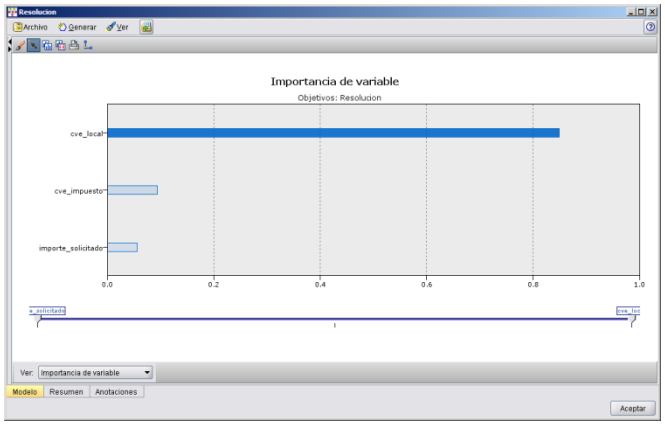


Figura IV.33. Importancia de las variables para la Red bayesiana.

Se puede observar que la clave de la local es la variable independiente más importante para el modelo.

La ruta final que generó todos estos resultados es la siguiente:

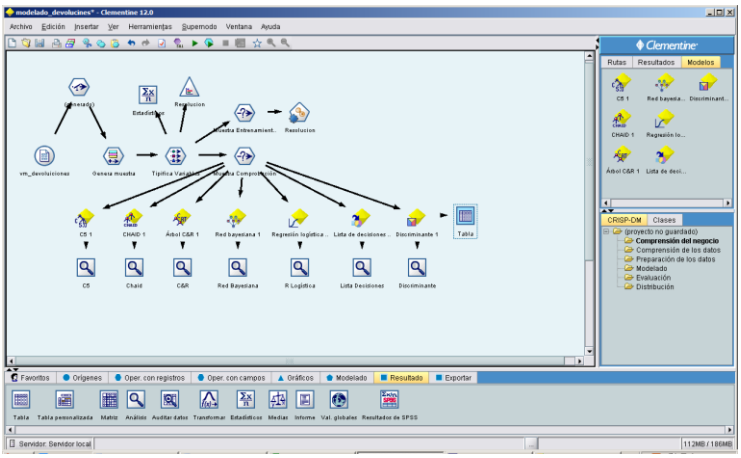


Figura IV.34. Ruta final en Clementine que generó el modelado para devoluciones.

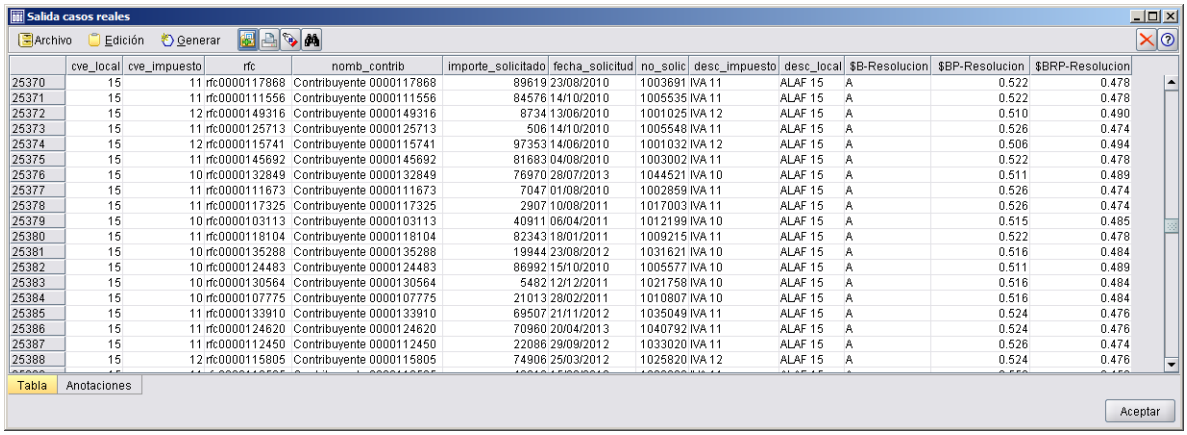
Para observar cómo se ven los resultados en una tabla de salida se muestra el siguiente ejemplo de los primeros registros para la muestra de comprobación:

cve_local	cve_impuesto	no_solic	porte_autoriza	cve_estatus	rfc	nomb_contrib	nporte_solicita	dias_transcurrid	desc_impuesto	desc_local	Partición	Resolucion	$B-Resolucion	$BP-Resolucion	$BRP-Resolucion
10	11	1008087	-	2	rfc0000143871	Contribuyente (	45,692	14	IVA 11	ALAF 10	2_Comprobacio	R	A	0.515638852	0.484361148
10	10	1000145	-	2	rfc0000112264	Contribuyente (	35,178	13	IVA 10	ALAF 10	2_Comprobacio	R	A	0.508033713	0.491966287
10	12	1012547	119,272	3	rfc0000122498	Contribuyente (	119,272	11	IVA 12	ALAF 10	2_Comprobacio	A	R	0.513631473	0.513631473
10	12	1021663	112,705	3	rfc0000105154	Contribuyente (	112,705	9	IVA 12	ALAF 10	2_Comprobacio	A	R	0.513631473	0.513631473
10	11	1008146	22,787	3	rfc0000119606	Contribuyente (	22,787	6	IVA 11	ALAF 10	2_Comprobacio	A	R	0.508576014	0.508576014
10	11	1036876	60,324	3	rfc0000128490	Contribuyente (	60,324	8	IVA 11	ALAF 10	2_Comprobacio	A	R	0.511053029	0.511053029
10	10	1000157	-	2	rfc0000128692	Contribuyente (	80,668	14	IVA 10	ALAF 10	2_Comprobacio	R	A	0.503232089	0.496767911
10	10	1000208	-	2	rfc0000140175	Contribuyente (	11,241	8	IVA 10	ALAF 10	2_Comprobacio	R	A	0.50898521	0.49101479
10	12	1021843	15,440	3	rfc0000124629	Contribuyente (	15,440	5	IVA 12	ALAF 10	2_Comprobacio	A	R	0.508858811	0.508858811
10	12	1042393	115,278	3	rfc0000146304	Contribuyente (	115,278	5	IVA 12	ALAF 10	2_Comprobacio	A	R	0.513631473	0.513631473
10	10	1002565	-	2	rfc0000144204	Contribuyente (	55,992	12	IVA 10	ALAF 10	2_Comprobacio	R	R	0.528486514	0.528486514
10	10	1034892	115,809	3	rfc0000113918	Contribuyente (	115,809	12	IVA 10	ALAF 10	2_Comprobacio	A	A	0.528817333	0.471182667
10	10	1032176	23,178	3	rfc0000145075	Contribuyente (	23,178	13	IVA 10	ALAF 10	2_Comprobacio	A	A	0.50898521	0.49101479
10	11	1015883	-	2	rfc0000122616	Contribuyente (	32,622	13	IVA 11	ALAF 10	2_Comprobacio	R	A	0.515638852	0.484361148
10	12	1024845	14,272	3	rfc0000100975	Contribuyente (	66,612	12	IVA 12	ALAF 10	2_Comprobacio	A	R	0.500898128	0.500898128
10	11	1008260	-	2	rfc0000112890	Contribuyente (	1,856	5	IVA 11	ALAF 10	2_Comprobacio	R	R	0.508576014	0.508576014
10	10	1020974	-	2	rfc0000141193	Contribuyente (	113,856	12	IVA 10	ALAF 10	2_Comprobacio	R	A	0.528817333	0.471182667
10	10	1034903	-	2	rfc0000126214	Contribuyente (	79,001	7	IVA 10	ALAF 10	2_Comprobacio	R	A	0.503232089	0.496767911
10	11	1019927	-	2	rfc0000113189	Contribuyente (	30,003	5	IVA 11	ALAF 10	2_Comprobacio	R	A	0.515638852	0.484361148
10	12	1026104	74,645	3	rfc0000133251	Contribuyente (	74,645	5	IVA 12	ALAF 10	2_Comprobacio	A	A	0.504629591	0.495370409
10	10	1000337	-	2	rfc0000116098	Contribuyente (	57,025	9	IVA 10	ALAF 10	2_Comprobacio	R	R	0.528486514	0.528486514
10	12	1010321	54,781	3	rfc0000123821	Contribuyente (	54,781	7	IVA 12	ALAF 10	2_Comprobacio	A	R	0.500898128	0.500898128
10	12	1000139	95,477	3	rfc0000102694	Contribuyente (	95,477	13	IVA 12	ALAF 10	2_Comprobacio	A	A	0.504629591	0.495370409
10	12	1023543	42,518	3	rfc0000149658	Contribuyente (	42,518	7	IVA 12	ALAF 10	2_Comprobacio	A	A	0.52672735	0.47327265
10	10	1042161	44,870	3	rfc0000133113	Contribuyente (	44,870	8	IVA 10	ALAF 10	2_Comprobacio	A	A	0.508033713	0.491966287
10	12	1022694	-	2	rfc0000148163	Contribuyente (	62,970	12	IVA 12	ALAF 10	2_Comprobacio	R	R	0.500898128	0.500898128
10	12	1042326	72,390	3	rfc0000134927	Contribuyente (	72,390	8	IVA 12	ALAF 10	2_Comprobacio	A	A	0.504629591	0.495370409

Figura IV.35. Muestra del resultado final para la muestra de comprobación que genera el modelo Red bayesiana para cada caso.

Las últimas tres variables son las que genera el modelo, $B-Resolucion es la variable que se pronostica, $BP-Resolucion es la estimación de probabilidad de la variable pronosticada y $BRP-Resolucion es el complemento de $BP cuando $B es igual a A (autorizado), Resolución es la variable objetivo observada, es decir la historia que se va a comparar contra la predicción.

Para poder evaluar los casos reales, aquellos de los que aún se desconoce su resultado se tendría que ejecutar el modelo ganador sobre estos casos sin evaluar para obtener la salida.



	cve_local	cve_impuesto	rfc	nomb_contrib	importe_solicitado	fecha_solicitud	no_solic	desc_impuesto	desc_local	$B-Resolucion	$BP-Resolucion	$BRP-Resolucion
25370	15	11	rfc0000117868	Contribuyente 0000117868	89619	23/08/2010	1003691	IVA 11	ALAF 15	A	0.522	0.478
25371	15	11	rfc0000111556	Contribuyente 0000111556	84576	14/10/2010	1005535	IVA 11	ALAF 15	A	0.522	0.478
25372	15	12	rfc0000149316	Contribuyente 0000149316	8734	13/06/2010	1001025	IVA 12	ALAF 15	A	0.510	0.490
25373	15	11	rfc0000125713	Contribuyente 0000125713	506	14/10/2010	1005548	IVA 11	ALAF 15	A	0.526	0.474
25374	15	12	rfc0000115741	Contribuyente 0000115741	97353	14/06/2010	1001032	IVA 12	ALAF 15	A	0.506	0.494
25375	15	11	rfc0000145692	Contribuyente 0000145692	81683	04/08/2010	1003002	IVA 11	ALAF 15	A	0.522	0.478
25376	15	10	rfc0000132849	Contribuyente 0000132849	76970	28/07/2013	1044521	IVA 10	ALAF 15	A	0.511	0.489
25377	15	11	rfc0000111673	Contribuyente 0000111673	7047	01/08/2010	1002859	IVA 11	ALAF 15	A	0.526	0.474
25378	15	11	rfc0000117325	Contribuyente 0000117325	2907	10/08/2011	1017003	IVA 11	ALAF 15	A	0.526	0.474
25379	15	10	rfc0000103113	Contribuyente 0000103113	40911	06/04/2011	1012199	IVA 10	ALAF 15	A	0.515	0.485
25380	15	11	rfc0000118104	Contribuyente 0000118104	82343	18/01/2011	1009215	IVA 11	ALAF 15	A	0.522	0.478
25381	15	10	rfc0000135288	Contribuyente 0000135288	19944	23/08/2012	1031621	IVA 10	ALAF 15	A	0.516	0.484
25382	15	10	rfc0000124483	Contribuyente 0000124483	86992	15/10/2010	1005577	IVA 10	ALAF 15	A	0.511	0.489
25383	15	10	rfc0000130564	Contribuyente 0000130564	5482	12/12/2011	1021758	IVA 10	ALAF 15	A	0.516	0.484
25384	15	10	rfc0000107775	Contribuyente 0000107775	21013	28/02/2011	1010807	IVA 10	ALAF 15	A	0.516	0.484
25385	15	11	rfc0000133910	Contribuyente 0000133910	69507	21/11/2012	1035049	IVA 11	ALAF 15	A	0.524	0.476
25386	15	11	rfc0000124620	Contribuyente 0000124620	70960	20/04/2013	1040792	IVA 11	ALAF 15	A	0.524	0.476
25387	15	11	rfc0000112450	Contribuyente 0000112450	22086	29/09/2012	1033020	IVA 11	ALAF 15	A	0.526	0.474
25388	15	12	rfc0000115805	Contribuyente 0000115805	74906	25/03/2012	1025820	IVA 12	ALAF 15	A	0.524	0.476

Figura IV.36. Muestra del resultado final para casos reales que genera el modelo Red bayesiana.

Se puede observar que el modelo solo necesita las variables independientes sin tener definida una variable dependiente, esto porque ésta será la $B-Resolución que es la predicción del modelo.

Este tipo de resultados se puede aplicar para muchos registros o uno solo, esto hace que una implementación pueda realizarse en un sistema que acepte la captura de las variables independientes o la carga masiva mediante archivos.

La ruta para ejecutar el modelo con datos nuevos se muestra a continuación.

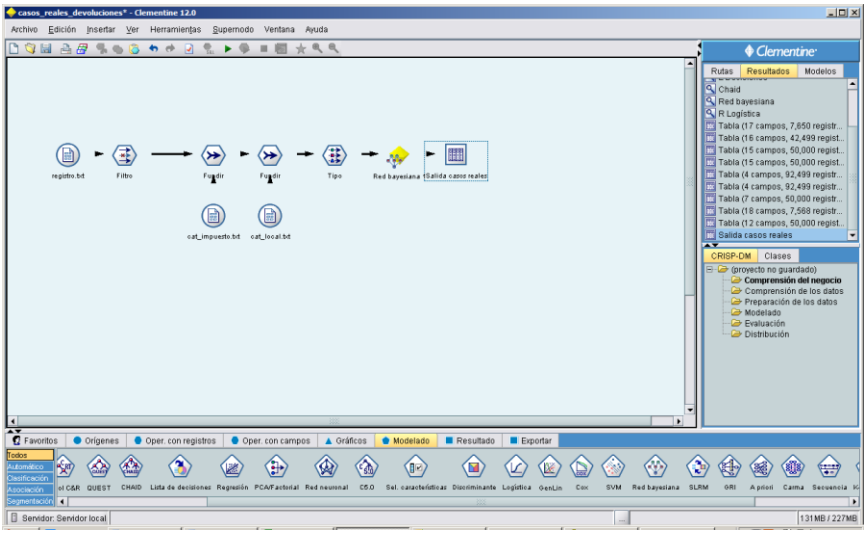


Figura IV.37. Ruta en Clementine para ejecutar el modelo con datos nuevos.

El nodo en amarillo es el modelo ganador que tiene todo el algoritmo necesario para calificar los nuevos casos que se necesiten.

La implementación de este tipo de modelos es sencilla ya que se pueden exportar a PMML (Predictive Model Markup Language), código XML que puede ser integrado en lenguajes de programación como Java, un extracto del modelo de Red bayesiana que acabamos de ver sería el siguiente:

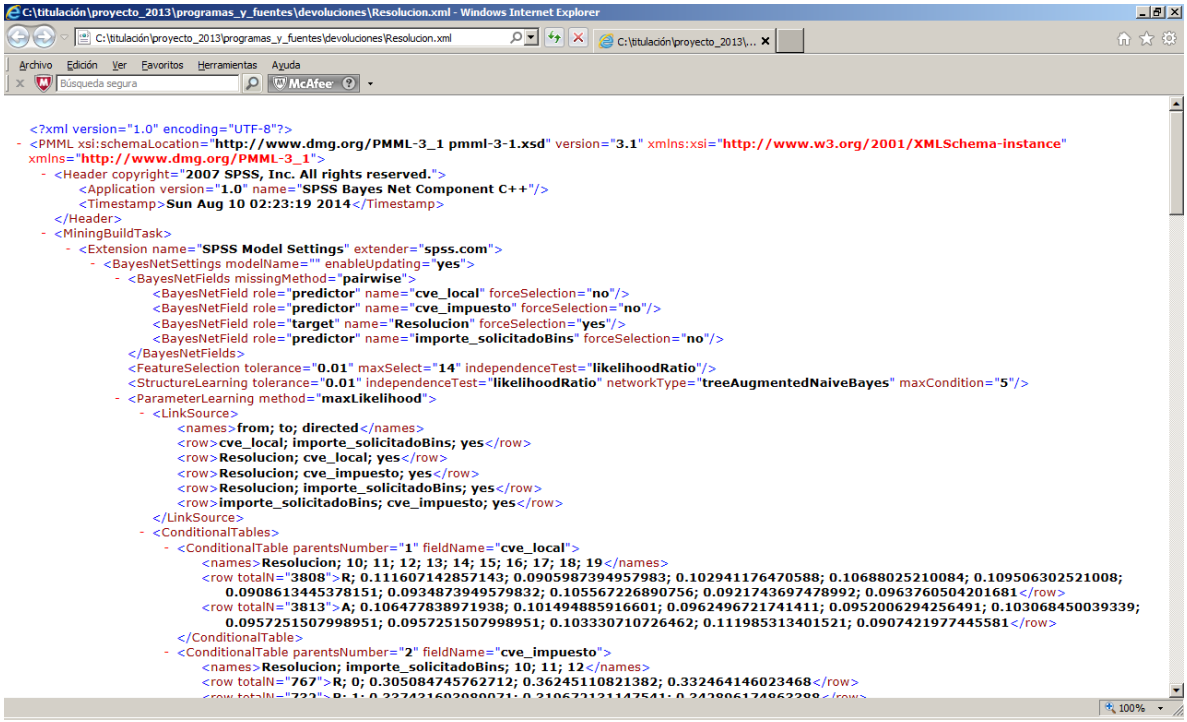


Figura IV.38. Muestra del código PMML que se genera al exportar el modelo Red bayesiana.

IV.2. Cobranza.

Como ya se vio para el modelado de devoluciones se empezará por la tipificación de los datos, las variables que se construyeron para la vista minable son las siguientes:

Variables	Descripción
cve_impuesto	Clave de impuesto.
cve_local	Clave de ALR.
no_credito	Número de crédito.
cve_baja	Clave de baja.
fecha_etapa	Fecha de la etapa en que se encuentra el crédito.
numero_etapas	Número de etapas por las que ha pasado un crédito.
rfc	Registro Federal de Contribuyentes.
nomb_contrib	Nombre del contribuyente.
importe	Importe del crédito fiscal.
fecha_origen	Fecha en que se originó el crédito.
fecha_registro	Fecha de registro en el área.
Pago	Definiciónde pago del crédito.
días_transc_origen	Días transcurridos desde el origen.
días_transc_registro	Días transcurridos desde el registro.
desc_local	Descripción de la ALR.
desc_impuesto	Descripción del impuesto.

Tabla IV.2. Campos que se utilizarán para el modelado en cobranza.

Recordemos que no todas las variables se ocuparán para el modelado pero si son importantes para identificar la calificación individual que el modelo le generará a cada contribuyente, las variables descriptivas se utilizan principalmente como salida final en consultas.

La ruta que muestra parte de estos datos es la siguiente:

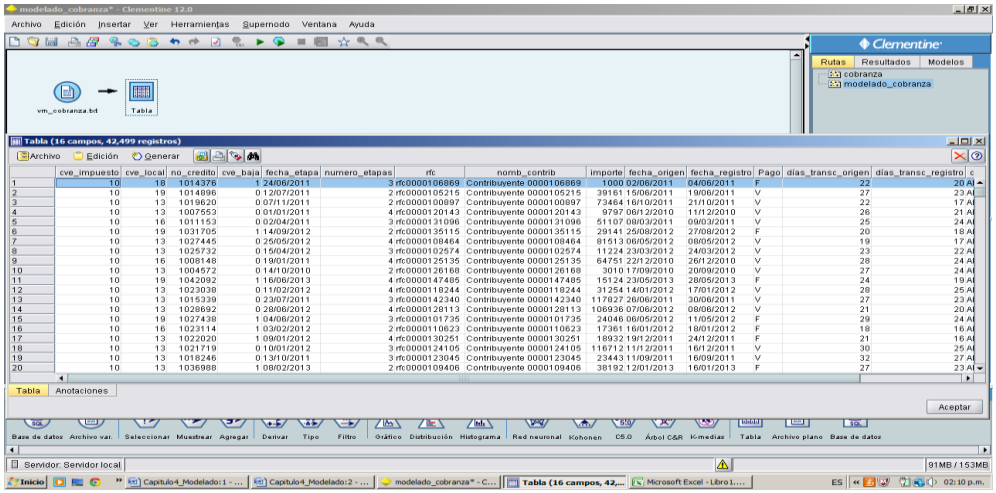


Figura IV.39. Ruta en Clementine que .genera una muestra de la información de la vista minable de Cobranza.

Añadiendo el nodo tipo a nuestra fuente de datos se tiene la siguiente información.

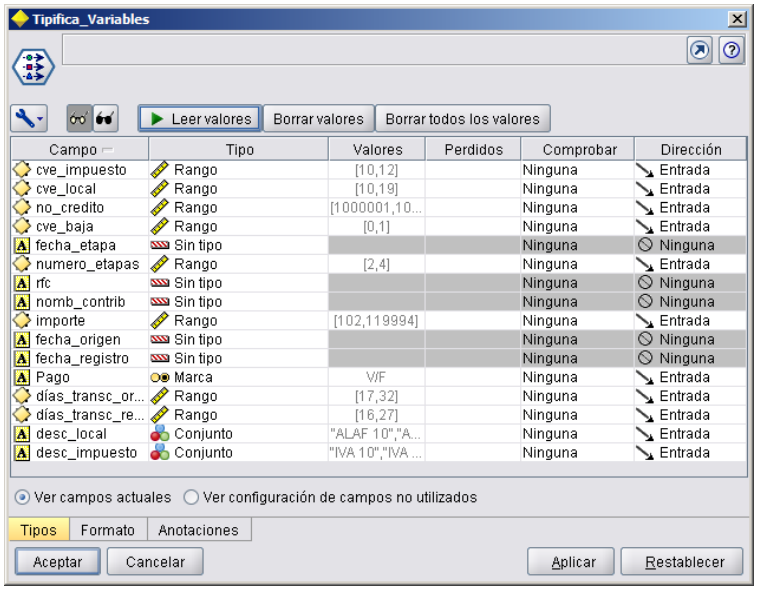


Figura IV.40. Nodo tipo con las propiedades de las variables de la vista minable de Cobranza.

Podemos observar que algunos tipos están mal definidos, esto lo corregimos como ya lo hemos visto anteriormente para tener las variables tipificadas correctamente y la dirección correcta quedando como sigue el nodo:

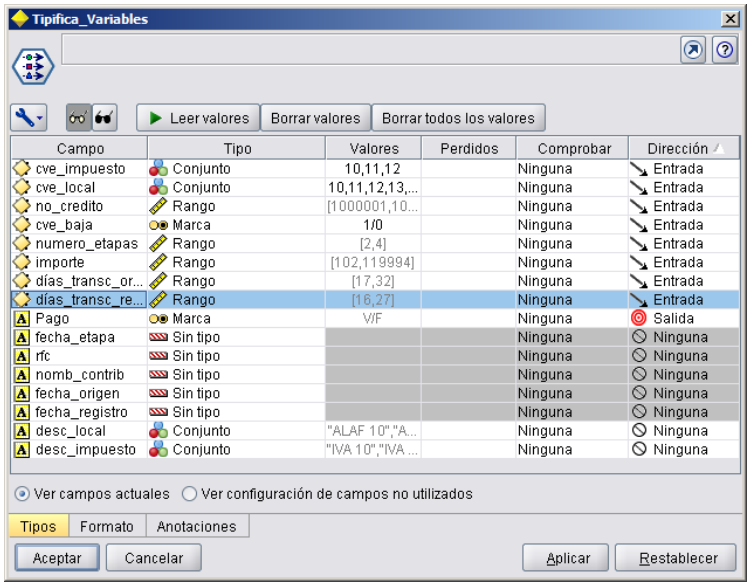


Figura IV.41. Corrección de los tipos de las variables y la entrada.

Se tienen potencialmente 8 variables de entrada y una de salida, recordemos que las descripciones no deben incluirse debido a que su clave ya está dentro de la selección y sería redundante, además que las claves son más fáciles de manejar por su simplicidad en diferentes modelos.

Para el modelo de devoluciones se utilizó una variable objetivo (salida) de tipo cadena con dos valores posibles que se tipificó como marca, para este modelo cambiaremos los valores de la variable de salida para que sean numéricos binario, 1 para pago (actualmente V) y 0 para no pago (F actualmente).

Esto lo realizamos con un nodo derivar de la siguiente manera:

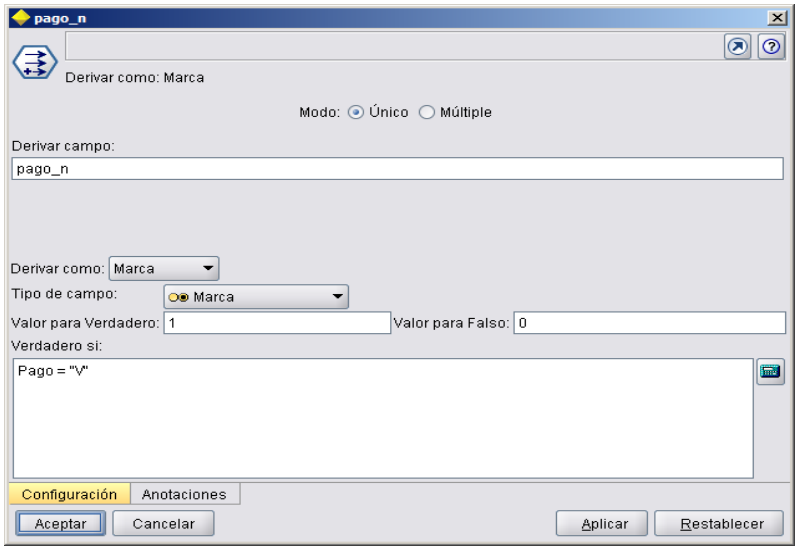


Figura IV.42. Creación de una nueva variable objetivo numérica a partir de la anterior de tipo cadena.

Podemos ver como queda la nueva variable objetivo, utilizando el nodo agrupar con la siguiente configuración:

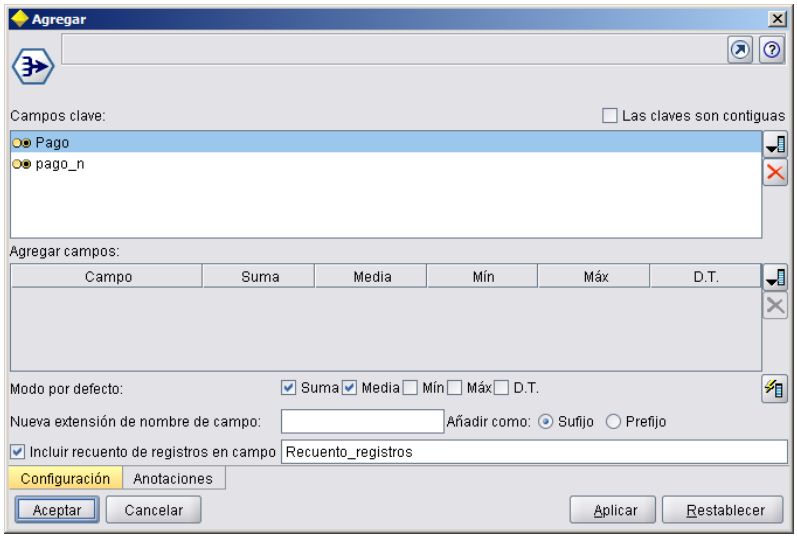


Figura IV.44. Configuración del nodo de Clementine Agregar para revisar los valores de la variable objetivo antigua y actual.

Esta configuración realizará una agrupación entre los valores de las dos variables y generará un conteo entre éstas, el resultado es el que se muestra a continuación:

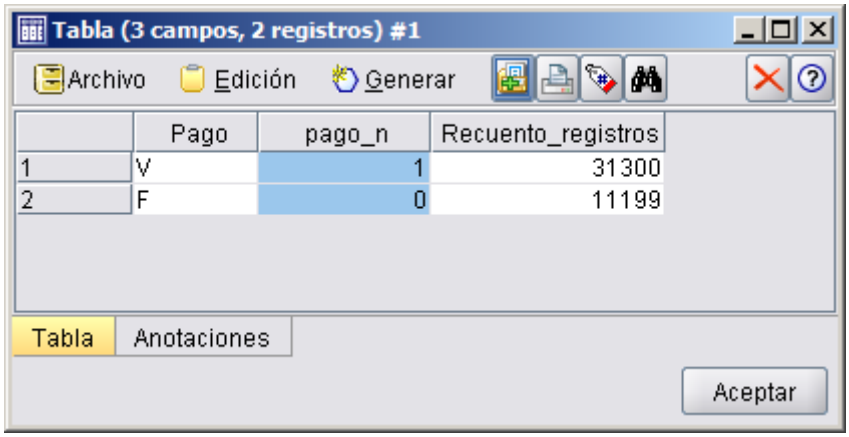


Figura IV.45. Resultado del nodo Agregar para la variable objetivo cadena (Pago) y para la nueva numérica (pago_n).

Podemos ver que como lo necesitamos ahora la variable pago_n tiene asignado un valor 1 cuando Pago es V (el crédito se pagó) y tiene 0 para el valor F (no se pagó el crédito fiscal). Con esta nueva redefinición la variable objetivo permitirá el uso de más modelos que trabajen con variables objetivo numéricas.

Como ya lo hemos realizado anteriormente es una buena práctica buscar variables correlacionadas para ir descartando y depurando la vista minable, utilizamos un nodo estadísticos con la siguiente configuración:

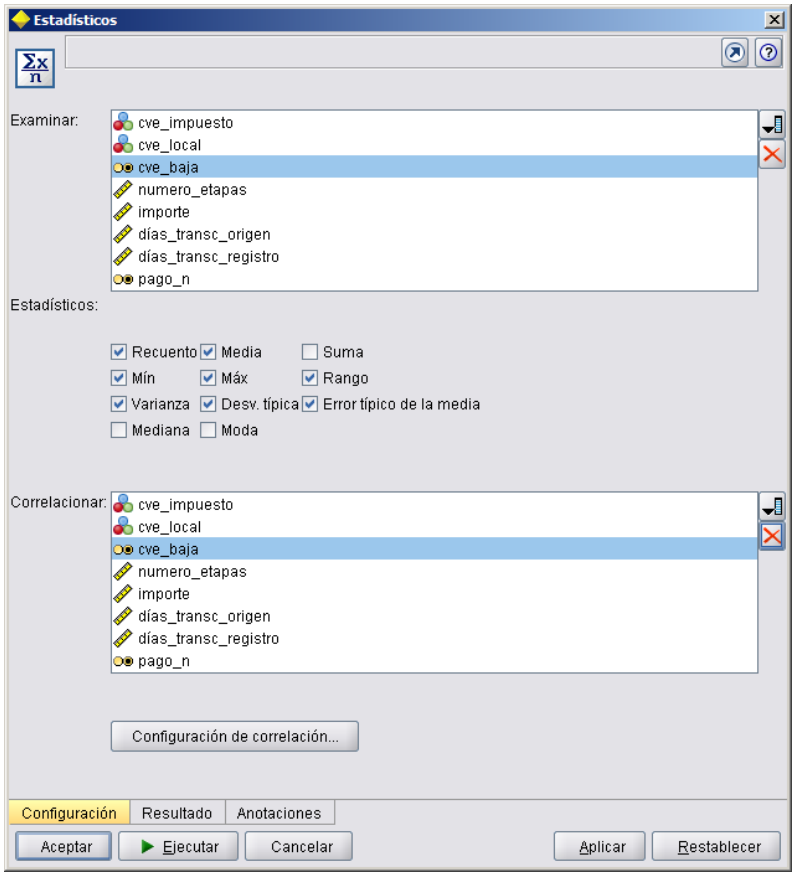


Figura IV.46. Configuración del nodo Estadísticos para encontrar correlaciones entre las variables independientes.

Los resultados de la correlación son:

Correlacciones de Pearson			
cve_impuesto	cve_local	0.001	Débil
	cve_baja	-0.001	Débil
	numero_etapas	0.006	Débil
	importe	0.008	Débil
	días_transc_origen	0.001	Débil
	días_transc_registro	0	Débil
	pago_n	0.001	Débil
cve_local	cve_impuesto	0.001	Débil
	cve_baja	-0.003	Débil
	numero_etapas	0.011	Fuerte
	importe	0.005	Débil
	días_transc_origen	0.003	Débil
	días_transc_registro	0.001	Débil
	pago_n	0.003	Débil
cve_baja	cve_impuesto	-0.001	Débil
	cve_local	-0.003	Débil
	numero_etapas	-0.334	Fuerte
	importe	0.006	Débil
	días_transc_origen	-0.004	Débil
	días_transc_registro	-0.002	Débil
	pago_n	-1	
numero_etapas	cve_impuesto	0.006	Débil
	cve_local	0.011	Fuerte
	cve_baja	-0.334	Fuerte
	importe	-0.001	Débil
	días_transc_origen	0.001	Débil
	días_transc_registro	-0.002	Débil
	pago_n	0.334	Fuerte
importe	cve_impuesto	0.008	Débil
	cve_local	0.005	Débil
	cve_baja	0.006	Débil
	numero_etapas	-0.001	Débil
	días_transc_origen	0.002	Débil
	días_transc_registro	0.001	Débil
	pago_n	-0.006	Débil
días_transc_origen	cve_impuesto	0.001	Débil
	cve_local	0.003	Débil
	cve_baja	-0.004	Débil
	numero_etapas	0.001	Débil
	importe	0.002	Débil
	días_transc_registro	0.925	Fuerte
	pago_n	0.004	Débil
días_transc_registro	cve_impuesto	0	Débil
	cve_local	0.001	Débil
	cve_baja	-0.002	Débil
	numero_etapas	-0.002	Débil
	importe	0.001	Débil
	días_transc_origen	0.925	Fuerte
	pago_n	0.002	Débil
pago_n	cve_impuesto	0.001	Débil
	cve_local	0.003	Débil
	cve_baja	-1	
	numero_etapas	0.334	Fuerte
	importe	-0.006	Débil
	días_transc_origen	0.004	Débil
	días_transc_registro	0.002	Débil

Figura IV.47. Resultados de la correlación para las variables independientes.

Se puede observar una correlación perfecta entre la variable objetivo y cve_baja, esto es normal y debimos haberlo detectado desde la preparación de los datos ya que cve_baja indica si el crédito se pagó o no, por lo que deberá ser excluida del modelado.

Por otro lado vemos una correlación muy fuerte entre los días transcurridos desde el origen del crédito y desde el registro, por lo que solo se deberá incluir una de las dos variables para el modelo.

Las otras correlaciones fuertes (excluyendo cve_baja) entre las variables son:

Variable 1	Variable 2	Explicación y diagnóstico
Cve_local	Numero_etapas	La correlación no es tan significativa, en ocasiones las variables parecieran tener correlación, pero al conocer el contenido de las variables y su fuente sabemos que es una coincidencia y no tiene importancia real este resultado, las variables seguirán en el modelo.
Pago_n	Numero_etapas	Esta es una correlación que conociendo el negocio, es normal, entre más tiempo pasa, es decir existen más número de etapas un crédito fiscal generalmente no se paga, se examinará al generar los modelos si se excluye la variable independiente
días_transc_origen	días_transc_registro	También en la correlación de ambas variables es normal, ya que son variables que en muchas ocasiones tienen orígenes iguales, es decir la fecha de origen a veces es muy cercana a la del registro, la primera es cuando se origina el crédito por lo cual se tomará como la que entrará al modelo y se excluirá la segunda

Tabla IV.3.Explicación y diagnóstico de las correlaciones fuertes entre las variables.

La ruta en Clementine que llevamos hasta el momento se muestra a continuación:

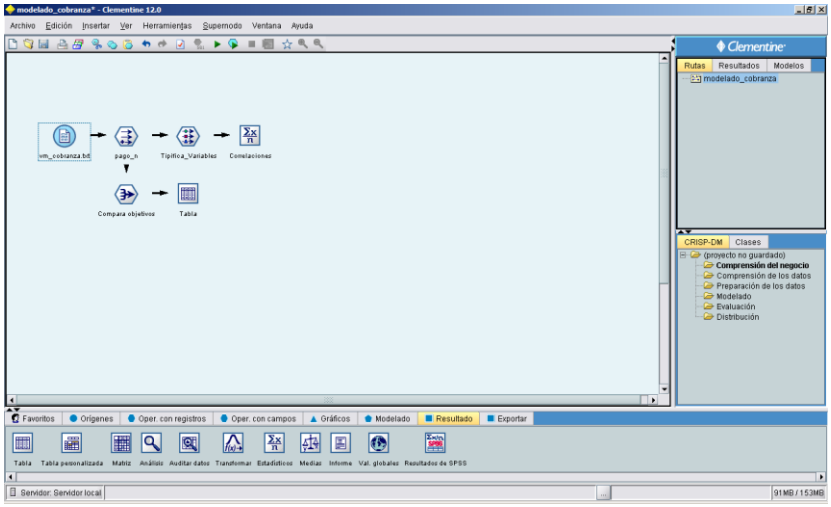


Figura IV.48. Ruta en Clementine con los pasos que se han realizado.

Como ya lo sabemos se tienen que generar las muestras de partición entrenamiento y comprobación, se añade el nodo correspondiente con el 70% y 30% respectivamente para cada una.

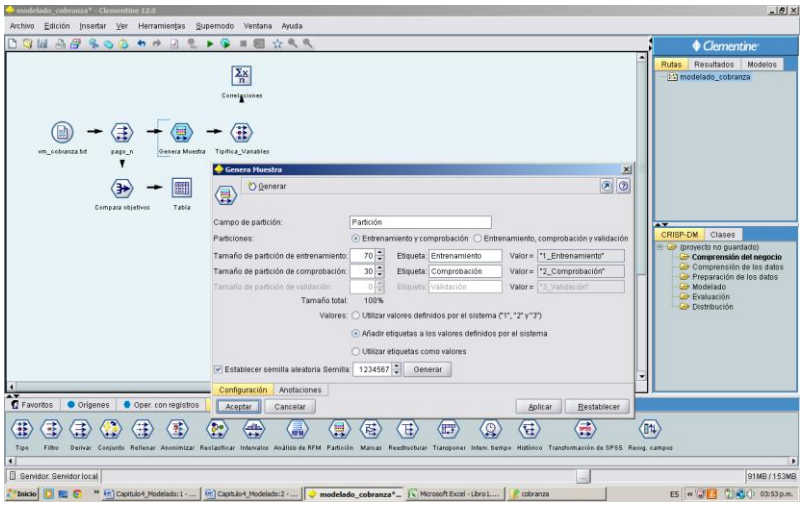


Figura IV.49. Ruta en Clementine que muestra los nodos y la configuración de la partición de entrenamiento y comprobación.

La configuración de variables de entrada y salida queda como sigue en el nodo tipo.

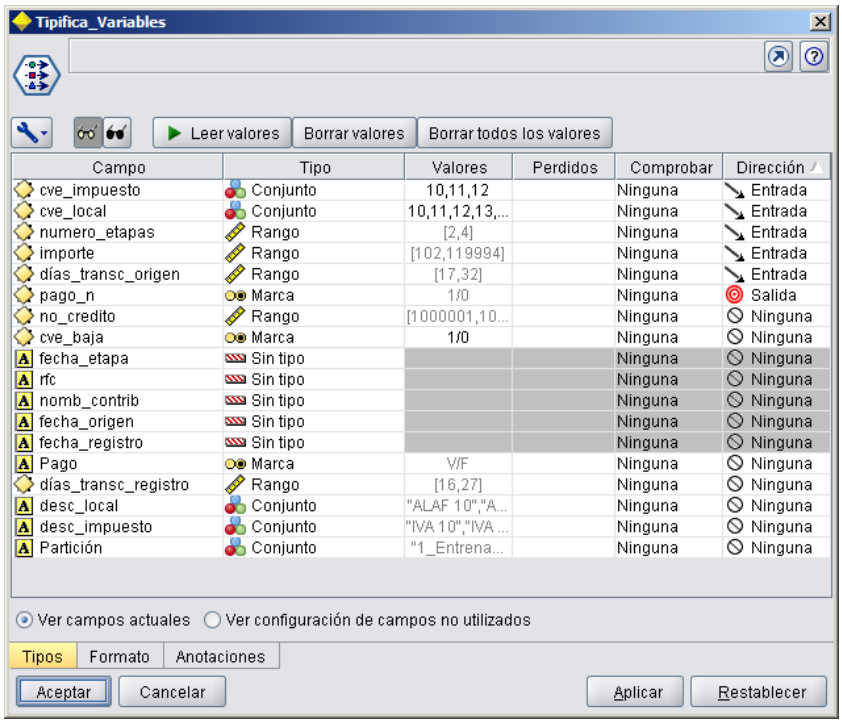


Figura IV.50. Configuración final para el nodo tipo.

5 variables independientes y la dependiente binaria ahora con valores numéricos.

Ya estamos en posibilidades de empezar con el modelado para los datos de cobranza, como ya lo hicimos anteriormente utilizaremos el clasificador binario ya que nuestra nueva variable objetivo sigue siendo de dos valores, el nodo podrá considerar ahora nuevos modelos que trabajan con variables objetivo numéricas.

Utilizando la muestra de entrenamiento la ruta se ve así:

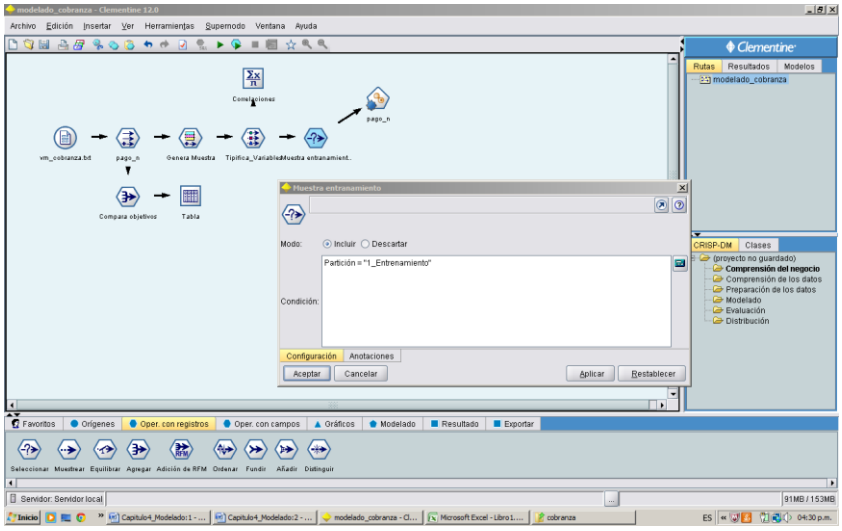
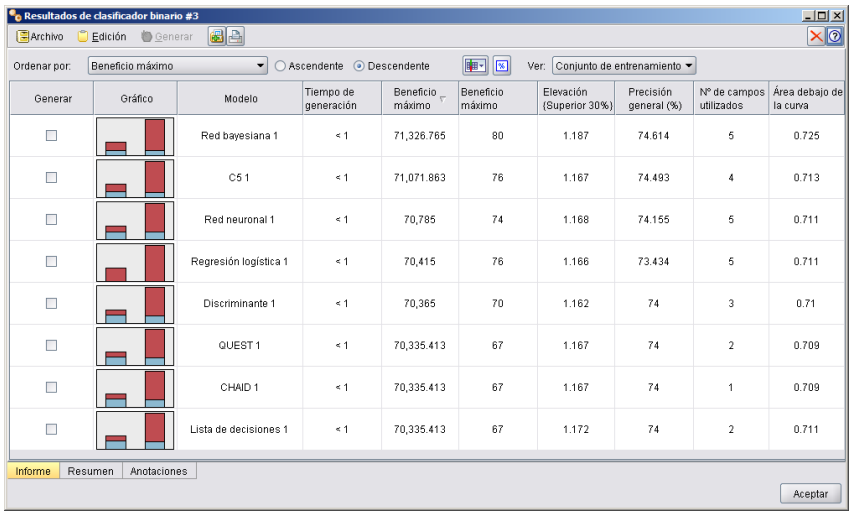


Figura IV.51. Ruta en Clementine con los nodos y se muestra el nodo seleccionar para la muestra de entrenamiento.

Veamos a continuación los resultados al ejecutar el clasificador binario:



Generar	Gráfico	Modelo	Tiempo de generación	Beneficio máximo	Beneficio máximo	Elevación (Superior 30%)	Precisión general (%)	Nº de campos utilizados	Área debajo de la curva
☐		Red bayesiana 1	< 1	71,326.765	80	1.187	74.614	5	0.725
☐		C5 1	< 1	71,071.863	76	1.167	74.493	4	0.713
☐		Red neuronal 1	< 1	70,785	74	1.168	74.155	5	0.711
☐		Regresión logística 1	< 1	70,415	76	1.166	73.434	5	0.711
☐		Discriminante 1	< 1	70,365	70	1.162	74	3	0.71
☐		QUEST 1	< 1	70,335.413	67	1.167	74	2	0.709
☐		CHAID 1	< 1	70,335.413	67	1.167	74	1	0.709
☐		Lista de decisiones 1	< 1	70,335.413	67	1.172	74	2	0.711

Figura IV.52. Resultados del clasificador binario para la generación de modelos.

Se muestran los resultados ordenados por el beneficio máximo de los modelos, para los fines de este trabajo se tomarán los primeros cinco modelos, éstos serán generados a partir de este momento para que puedan ser evaluados.

Se conectan al nodo seleccionar muestra de entrenamiento cada uno de los modelos y se deja la configuración por defecto de cada uno para generar los resultados como se ve a continuación en la ruta de Clementine.

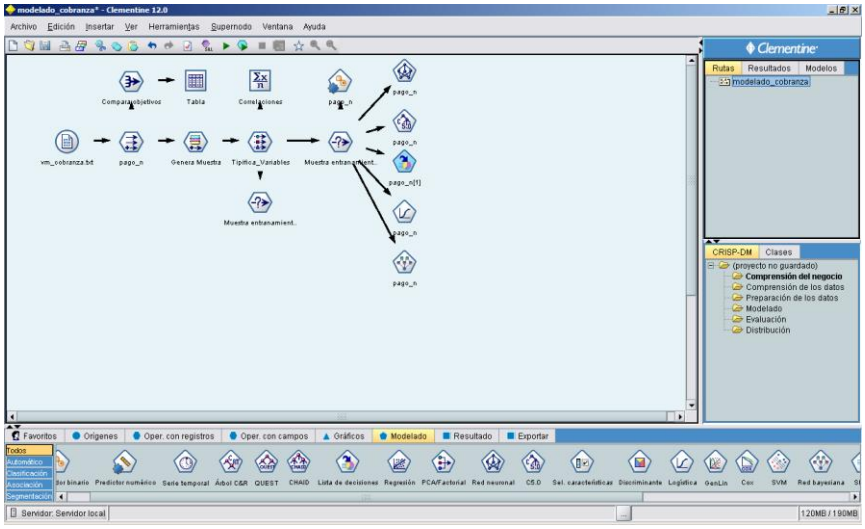


Figura IV.53. Ruta en Clementine que muestra los nodos creados y al final la conexión con los nodos de modelos a ejecutar.

Los cinco modelos se generan en la paleta de modelos del lado derecho del software de minería de datos:

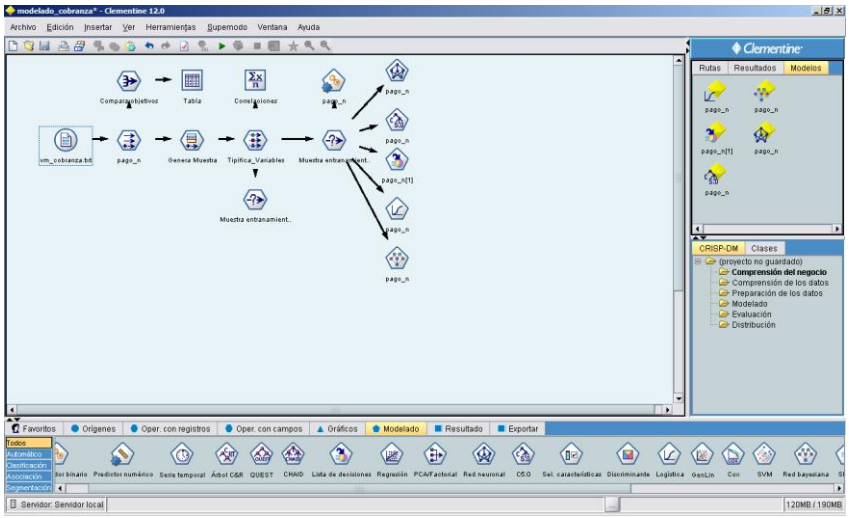


Figura IV.54. Ruta en Clementine que muestra la creación de los 5 modelos en la paleta de salida.

En cada uno de los modelos generados se puede observar en la pestaña resumen las variables que ocuparon para la generación, a continuación como ejemplo se muestra la de la regresión lineal logística:

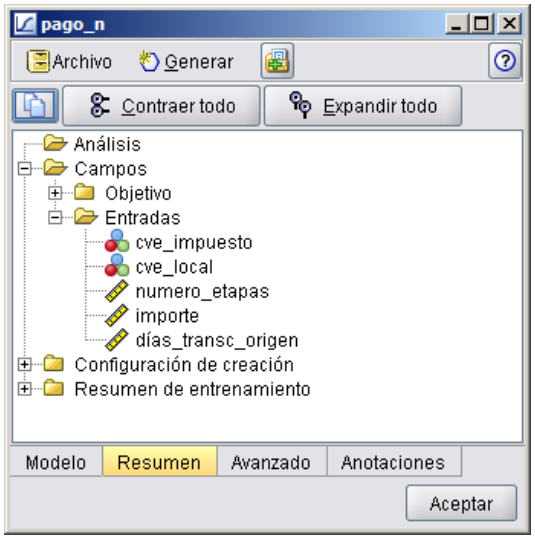


Figura IV.55. Variables ocupadas por el modelo Regresión lineal en la pestaña resumen.

La siguiente tabla muestra para cada uno las variables ocupadas y la importancia de éstas (cuando el modelo las calcula:

Modelo	Variables	Importancia (1 = más importante)
Red Bayesiana	numero_etapas	1
	importe	2
	cve_impuesto	3
	cve_local	4
	dias_transc_origen	5
C5	numero_etapas	1
	cve_impuesto	2
	dias_transc_origen	3
	cve_local	4
Red Neuronal	numero_etapas	1
	cve_local	2
	importe	3
	dias_transc_origen	4
	cve_impuesto	5
Regresión Logística	numero_etapas	1
	importe	2
	cve_local	3
	cve_impuesto	4
	dias_transc_origen	5
Discriminante	dias_transc_origen	No disponible
	numero_etapas	

Figura IV.56. Variables utilizadas y su importancia por cada modelo.

El modelo que menos variables ocupa es el de Discriminante con dos, el árbol C5 ocupa 4, todos los demás ocupan las 5 variables de entrada.

Veamos a continuación el desempeño de cada uno con la muestra de comprobación. Para esto utilizaremos el nodo analizar como ya lo vimos para devoluciones.

El cuadro siguiente muestra la evaluación de los casos erróneos contra los correctos para la muestra de comprobación:

Resultados de los modelos, muestra de comprobación										
Clasificación	Análisis Discriminante		Regresión Logística		Red Bayesiana		C5		Red Neuronal	
	Casos	%	Casos	%	Casos	%	Casos	%	Casos	%
Correctos	7,403	57.67%	9,518	74.15%	9,471	73.78%	9,491	73.93%	9,483	73.87%
Erróneos	5,434	42.33%	3,319	25.85%	3,366	26.22%	3,346	26.07%	3,354	26.13%
Total Casos	12,837									

Figura IV.57. Resumen del resultado de los modelos para la muestra de comprobación.

Podemos observar que el modelo que mejor desempeño tiene en la clasificación de nuevos casos es la regresión logística, mientras que el peor es al análisis de discriminante.

Esto nos puede indicar el mejor modelo para pronosticar correctamente nuevos casos que no han sido pronosticados.

La ruta final que genera todos los resultados vistos es la siguiente:

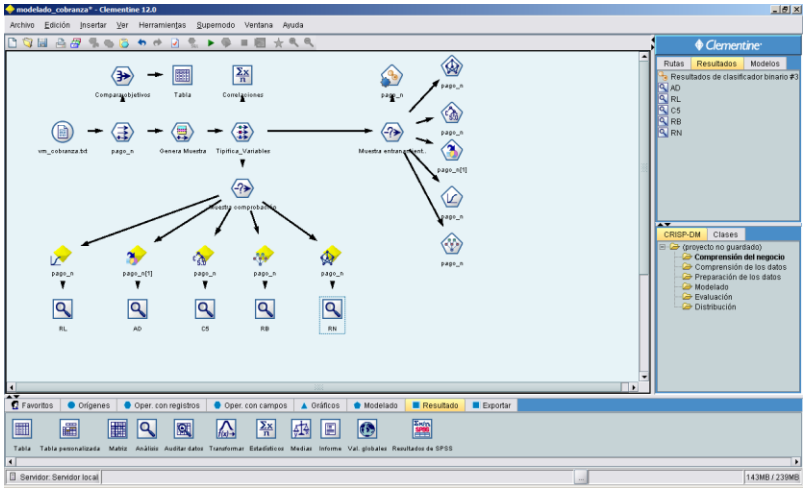


Figura IV.58. Ruta final con los pasos para la creación de los modelos.

Tomemos los dos modelos más precisos para revisar más a fondo sus detalles en los resultados generados:

1. ***Regresión Logística.***

En la pestaña modelo de este algoritmo muestra la siguiente información:

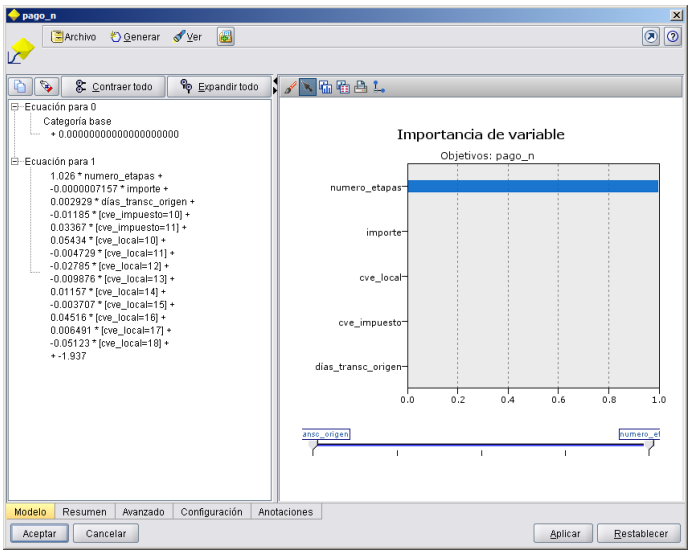


Figura IV.59. Resultados del modelo Regresión logística, muestra las ecuaciones y la importancia de las variables.

Por un lado vemos las ecuaciones generadas y por el otro la importancia de las variables.

En la pestaña avanzado vemos los resultados de la regresión:

Resumen del procesamiento de los casos

		N	Porcentaje marginal
pago_n	0	7880	26.60%
	1	21782	73.40%
cve_impuesto	10	9796	33.00%
	11	10000	33.70%
	12	9866	33.30%
cve_local	10	3006	10.10%
	11	2971	10.00%
	12	2933	9.90%
	13	2879	9.70%
	14	3002	10.10%
	15	3028	10.20%
	16	2936	9.90%
	17	2943	9.90%
	18	2959	10.00%
	19	3005	10.10%
Válidos		29662	100.00%
Perdidos		0	
Total		29662	
Subpoblación		29656(a)	

a. La variable dependiente sólo tiene un valor observado en 29654 (100.0%) subpoblaciones.

Información del ajuste del modelo

	Criterio de ajuste del modelo	Contrastes de la razón de verosimilitud		
Modelo	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersección	34339.535			
Final	30811.404	3528.13	14	0

Pseudo R-cuadrado

Cox y Snell	0.112
Nagelkerke	0.164
McFadden	0.103

Estimaciones de los parámetros

pago_n(a)		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
								Límite inferior	Límite superior
1	Intersección	-1.937	0.118	268.813	1	0			
	numero_etapas	1.026	0.019	3022.921	1	0	2.791	2.691	2.895
	importe	0	0	3.129	1	0.077	1	1	1
	días_transc_origen	0.003	0.004	0.61	1	0.435	1.003	0.996	1.01
	[cve_impuesto=10]	-0.012	0.034	0.12	1	0.729	0.988	0.924	1.057
	[cve_impuesto=11]	0.034	0.034	0.966	1	0.326	1.034	0.967	1.106
	[cve_impuesto=12]	0(b)	.	.	0	.	.	.	.
	[cve_local=10]	0.054	0.063	0.754	1	0.385	1.056	0.934	1.194
	[cve_local=11]	-0.005	0.062	0.006	1	0.939	0.995	0.881	1.124
	[cve_local=12]	-0.028	0.062	0.199	1	0.655	0.973	0.861	1.099
	[cve_local=13]	-0.01	0.063	0.025	1	0.875	0.99	0.875	1.12
	[cve_local=14]	0.012	0.062	0.034	1	0.853	1.012	0.895	1.143
	[cve_local=15]	-0.004	0.062	0.004	1	0.952	0.996	0.882	1.125
	[cve_local=16]	0.045	0.063	0.514	1	0.474	1.046	0.925	1.184
	[cve_local=17]	0.006	0.063	0.011	1	0.917	1.007	0.89	1.138
	[cve_local=18]	-0.051	0.062	0.676	1	0.411	0.95	0.841	1.073
	[cve_local=19]	0(b)	.	.	0	.	.	.	.

a. La categoría de referencia es: 0.

b. Este parámetro se ha establecido a cero porque es redundante.

Figura IV.60. Resultados del modelo Regresión logística.

2. Árbol de decisión C5.

La pestaña modelo de este algoritmo es la siguiente:

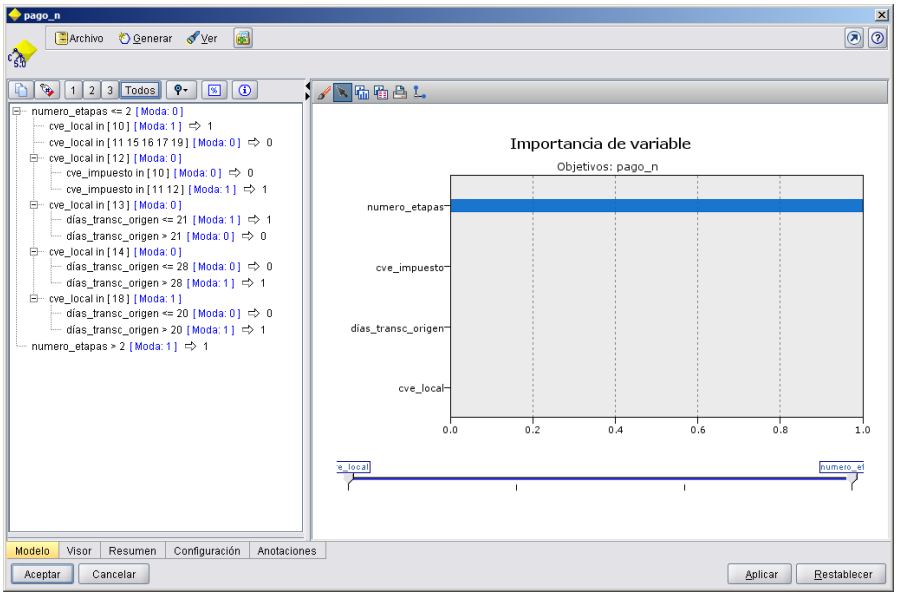


Figura IV.61. Resultados del modelo Árbol de decisión C5 con las condiciones y la importancia de las variables.

Por un lado se muestra las condiciones que genera el árbol como código y por el otro la importancia de las variables.

Para la pestaña visor se muestra:

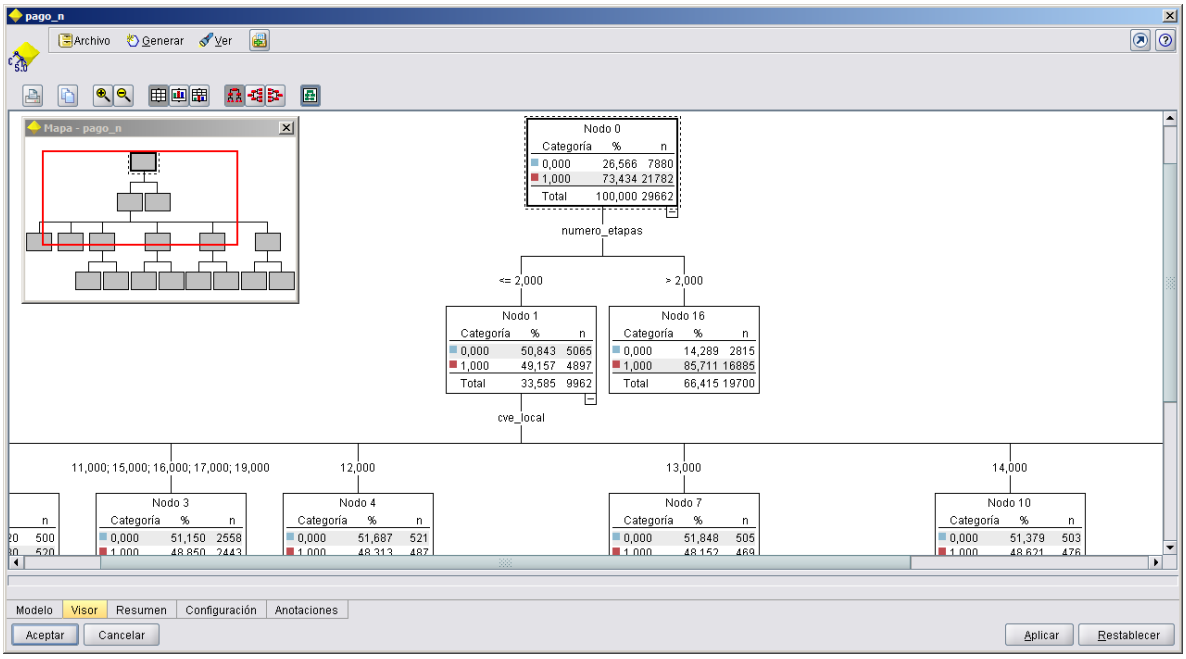


Figura IV.62. Visor del árbol de decisión C5 generado.

Se puede observar de forma gráfica el árbol generado por el algoritmo.

Podemos tener las siguientes conclusiones viendo los resultados que se mostraron arriba:

- El modelo de Regresión Logística aún siendo más preciso requiere de conocimientos matemáticos más especializados para interpretar los resultados.

- El modelo de árbol de decisión aún y cuando es menos preciso muestra resultados más intuitivos cuando se revisan los gráficos generados.

- Se puede ocupar el modelo más preciso para el pronóstico de casos nuevos en específico, ya que no se necesita una explicación mayor que un grado de estimación de probabilidad.

- Para generar estrategias que puedan ayudar en la cobranza se puede ver en el árbol de decisión que cuando el número de etapas es menor a 2 el porcentaje de cobro es de más

del 85% por lo cual es de vital importancia tener acciones tempranas para la recuperación de los adeudos.

Como vemos se pueden combinar los resultados de dos modelos predictivos para obtener diferentes beneficios a un mismo problema.

También ambos modelos pueden ser exportados como XML (PMML) para poder ser implementados dentro de soluciones que ya tenga la institución o crear a partir de estos herramientas informáticas que permitan la toma de decisiones.

Una vez con los dos modelos que podemos utilizar para implementar los casos nuevos la ruta que generaría los resultados es la siguiente:

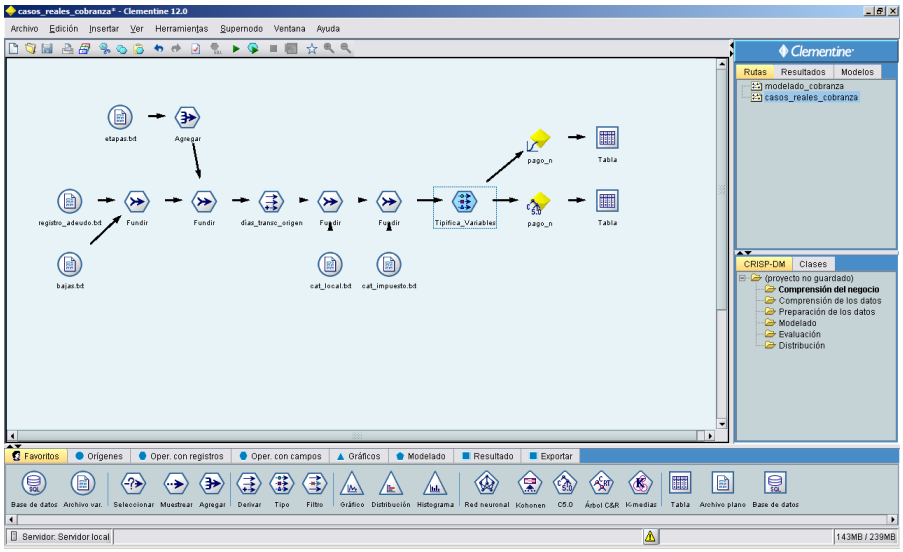
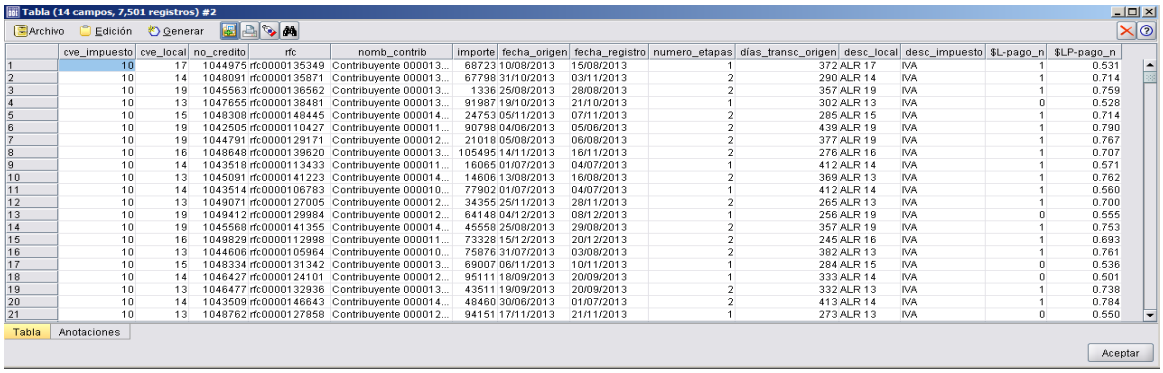


Figura IV.63. Ruta con los modelos generados.

Vemos que los dos modelos ganadores se conectan a los casos nuevos que no han tenido un final en el cobro de créditos fiscales, los resultados del modelo con mayor precisión se muestran a continuación:

Regresión logística:



	cve_impuesto	cve_local	no_credito	rfc	nomb_contrib	importe	fecha_origen	fecha_registro	numero_etapas	dias_transc_origen	desc_local	desc_impuesto	$L-pago_n	$LP-pago_n
1	10	17	1044975	rfc0000135349	Contribuyente 000013...	68723	10/08/2013	15/08/2013	1	372	ALR 17	IVA	1	0.531
2	10	14	1048091	rfc0000135871	Contribuyente 000013...	67798	31/10/2013	03/11/2013	2	290	ALR 14	IVA	1	0.714
3	10	19	1045563	rfc0000136562	Contribuyente 000013...	1336	25/08/2013	28/08/2013	2	357	ALR 19	IVA	1	0.759
4	10	13	1047655	rfc0000138481	Contribuyente 000013...	91987	19/10/2013	21/10/2013	1	302	ALR 13	IVA	0	0.528
5	10	15	1048308	rfc0000148445	Contribuyente 000014...	24753	05/11/2013	07/11/2013	2	285	ALR 15	IVA	1	0.714
6	10	19	1042505	rfc0000110427	Contribuyente 000011...	90798	04/06/2013	05/06/2013	2	439	ALR 19	IVA	1	0.790
7	10	19	1044791	rfc0000129171	Contribuyente 000012...	21018	05/08/2013	06/08/2013	2	377	ALR 19	IVA	1	0.767
8	10	16	1048648	rfc0000139620	Contribuyente 000013...	105495	14/11/2013	16/11/2013	2	276	ALR 16	IVA	1	0.707
9	10	14	1043518	rfc0000113433	Contribuyente 000011...	16065	01/07/2013	04/07/2013	1	412	ALR 14	IVA	1	0.571
10	10	13	1045091	rfc0000141223	Contribuyente 000014...	14606	13/08/2013	16/08/2013	2	369	ALR 13	IVA	1	0.762
11	10	14	1043514	rfc0000106783	Contribuyente 000010...	77902	01/07/2013	04/07/2013	1	412	ALR 14	IVA	1	0.560
12	10	13	1049071	rfc0000127005	Contribuyente 000012...	34355	25/11/2013	28/11/2013	2	265	ALR 13	IVA	1	0.700
13	10	19	1049412	rfc0000129984	Contribuyente 000012...	64148	04/12/2013	08/12/2013	1	256	ALR 19	IVA	0	0.555
14	10	19	1045568	rfc0000141355	Contribuyente 000014...	45558	25/08/2013	29/08/2013	2	357	ALR 19	IVA	1	0.753
15	10	16	1049829	rfc0000112998	Contribuyente 000011...	73328	15/12/2013	20/12/2013	2	245	ALR 16	IVA	1	0.693
16	10	13	1044606	rfc0000105964	Contribuyente 000010...	75876	31/07/2013	03/08/2013	2	382	ALR 13	IVA	1	0.761
17	10	15	1048334	rfc0000131342	Contribuyente 000013...	69007	06/11/2013	10/11/2013	1	284	ALR 15	IVA	0	0.536
18	10	14	1046427	rfc0000124101	Contribuyente 000012...	95111	18/09/2013	20/09/2013	1	333	ALR 14	IVA	0	0.501
19	10	13	1046477	rfc0000132936	Contribuyente 000013...	43511	19/09/2013	20/09/2013	2	332	ALR 13	IVA	1	0.738
20	10	14	1043509	rfc0000146643	Contribuyente 000014...	48460	30/06/2013	01/07/2013	2	413	ALR 14	IVA	1	0.784
21	10	13	1048762	rfc0000127858	Contribuyente 000012...	94151	17/11/2013	21/11/2013	1	273	ALR 13	IVA	0	0.550

Figura IV.64. Muestra con los resultados de los datos para la regresión logística.

Las dos últimas variables son las que generara el modelo como pronóstico para cada caso nuevo, por ejemplo para el primer registros el contribuyente con número de crédito 1044975 (el primero de la tabla anterior) el modelo predice que se pagará con una estimación de probabilidad de .53 el contribuyente con número de crédito 1043509 (de la fila 20) pagará el crédito con una estimación de probabilidad de .78.

Estos resultados pueden utilizarse para tener una estimación en la recuperación de importes individuales para después tener una estimación de recuperación global de la cartera. Veamos para cada crédito fiscal se puede aplicar la siguiente fórmula en un nodo derivar que llamaremos estimación de recuperación.

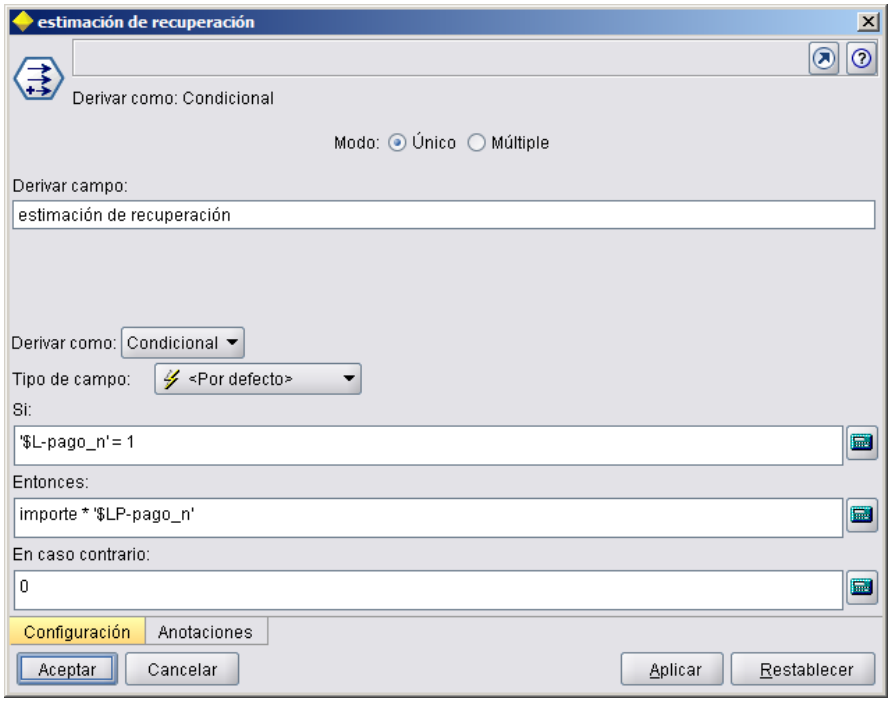
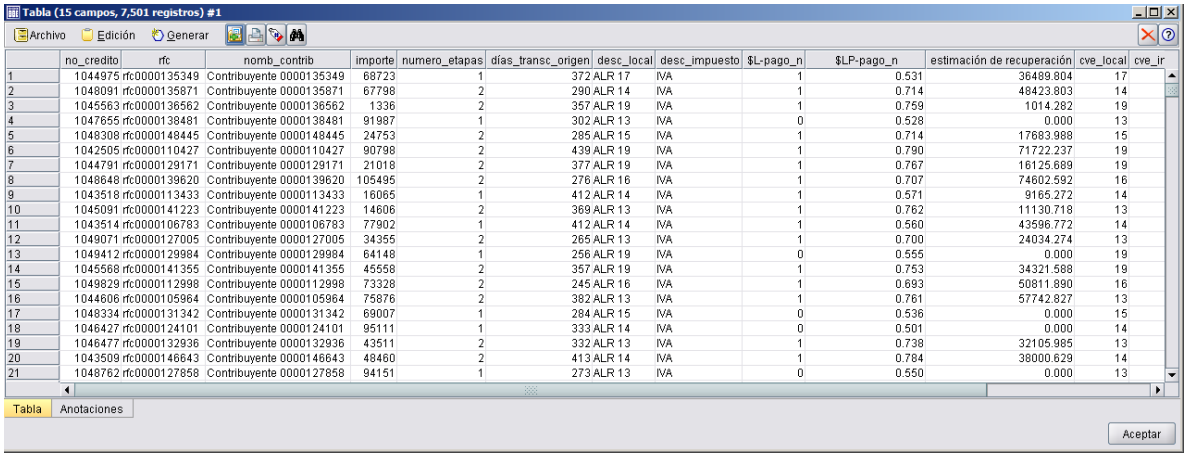


Figura IV.65. Nodo derivar para la creación de la variable estimación de recuperación.

Cuando el modelo estime que el crédito fiscal se va a pagar, entonces el importe a recuperar es de al menos el importe del crédito por la estimación de probabilidad.

Los resultados se pueden ver en la siguiente tabla:



	no_credito	rfc	nomb_contrib	importe	numero_etapas	días_transc_origen	desc_local	desc_impuesto	$L-pago_n	$LP-pago_n	estimación de recuperación	cve_local	cve_ir
1	1044975	rfc0000135349	Contribuyente 0000135349	68723	1	372	ALR 17	IVA	1	0.531	36489.804	17	
2	1048091	rfc0000135871	Contribuyente 0000135871	67798	2	290	ALR 14	IVA	1	0.714	48423.803	14	
3	1045563	rfc0000136562	Contribuyente 0000136562	1336	2	357	ALR 19	IVA	1	0.759	1014.282	19	
4	1047655	rfc0000138481	Contribuyente 0000138481	91987	1	302	ALR 13	IVA	0	0.528	0.000	13	
5	1048308	rfc0000148445	Contribuyente 0000148445	24753	2	285	ALR 15	IVA	1	0.714	17683.988	15	
6	1042505	rfc0000110427	Contribuyente 0000110427	90798	2	439	ALR 19	IVA	1	0.790	71722.237	19	
7	1044791	rfc0000129171	Contribuyente 0000129171	21018	2	377	ALR 19	IVA	1	0.767	16125.689	19	
8	1048648	rfc0000139620	Contribuyente 0000139620	105495	2	276	ALR 16	IVA	1	0.707	74602.592	16	
9	1043518	rfc0000113433	Contribuyente 0000113433	16065	1	412	ALR 14	IVA	1	0.571	9165.272	14	
10	1045091	rfc0000141223	Contribuyente 0000141223	14606	2	369	ALR 13	IVA	1	0.762	11130.718	13	
11	1043514	rfc0000106783	Contribuyente 0000106783	77902	1	412	ALR 14	IVA	1	0.560	43596.772	14	
12	1049071	rfc0000127005	Contribuyente 0000127005	34355	2	265	ALR 13	IVA	1	0.700	24034.274	13	
13	1049412	rfc0000129984	Contribuyente 0000129984	64148	1	256	ALR 19	IVA	0	0.555	0.000	19	
14	1045568	rfc0000141355	Contribuyente 0000141355	45558	2	357	ALR 19	IVA	1	0.753	34321.588	19	
15	1049829	rfc0000112998	Contribuyente 0000112998	73328	2	245	ALR 16	IVA	1	0.693	50811.890	16	
16	1044606	rfc0000105964	Contribuyente 0000105964	75876	2	382	ALR 13	IVA	1	0.761	57742.827	13	
17	1048334	rfc0000131342	Contribuyente 0000131342	69007	1	284	ALR 15	IVA	0	0.536	0.000	15	
18	1046427	rfc0000124101	Contribuyente 0000124101	95111	1	333	ALR 14	IVA	0	0.501	0.000	14	
19	1046477	rfc0000132936	Contribuyente 0000132936	43511	2	332	ALR 13	IVA	1	0.738	32105.985	13	
20	1043509	rfc0000146643	Contribuyente 0000146643	48460	2	413	ALR 14	IVA	1	0.784	38000.629	14	
21	1048762	rfc0000127858	Contribuyente 0000127858	94151	1	273	ALR 13	IVA	0	0.550	0.000	13	

Figura IV.66. Muestra de los resultados con la variable estimación de recuperación.

Se tiene una estimación de recuperación por cada contribuyente, si sumamos esta última columna creada para los 7,501 créditos y el importe de cada uno, utilizamos el nodo agregar para esta operación.

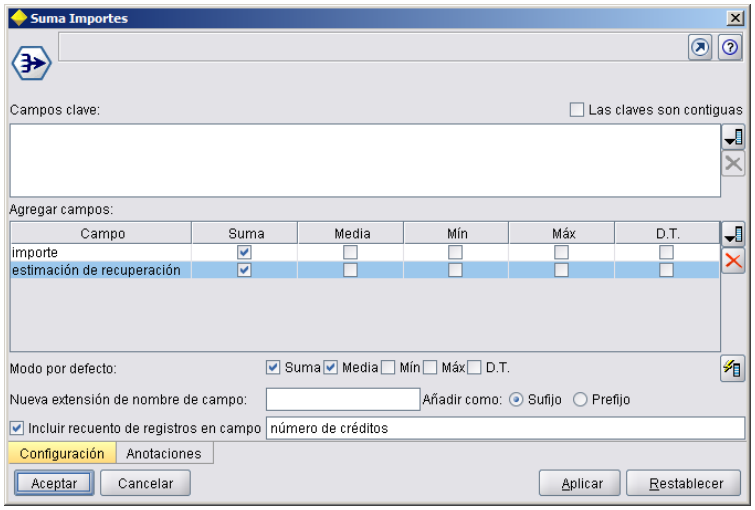


Figura IV.67. Configuración del nodo agregar para el importe y la estimación de recuperación.

Se obtiene el siguiente resultado:

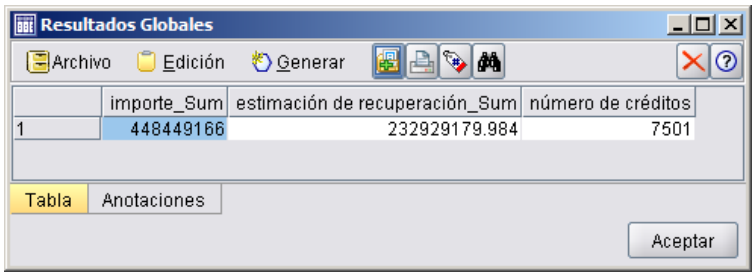


Figura IV.68. Resultados globales del importe, estimación de recuperación y el número de créditos.

Esto se puede interpretar de la siguiente manera.

Se tienen un total de 7,501 créditos fiscales en cartera activa, con un adeudo de $448,449,166, del cual se espera recuperar al menos $232,929,179.98, podríamos decir que ese es el valor esperado de la cartera. La ruta final que realiza estos cálculos se muestra a continuación:

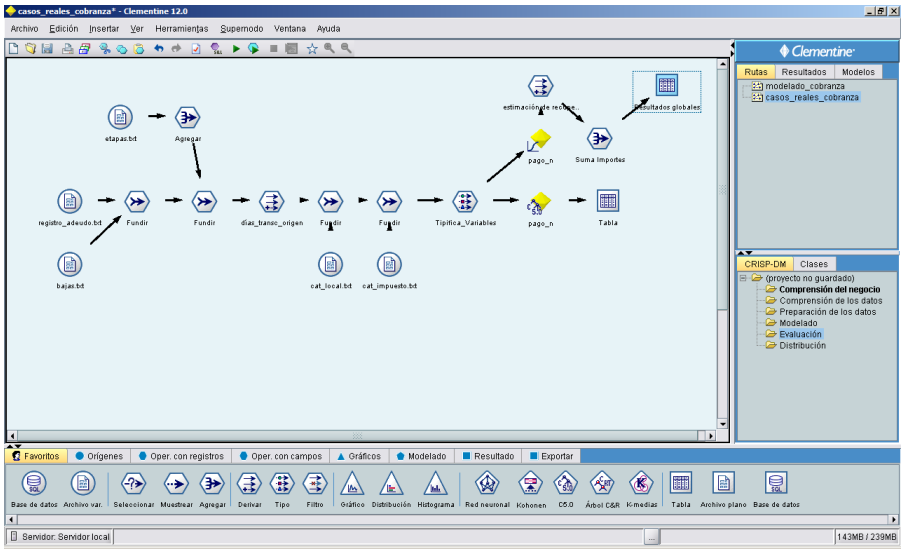


Figura IV.69. Ruta final para el modelado en Cobranza.

IV.3. Ciclo de vida de los modelos.

Por último es importante mencionar el ciclo de vida de los modelos, éstos no son eternos, debe de hacerse revisiones periódicas para analizar los resultados que genera y evaluar si aún está trabajando de forma correcta.

Se pueden tener indicadores que midan periódicamente la eficiencia en los resultados obtenidos (por ejemplo en la cobranza, revisar que el pronóstico de pago se cumpla en la realidad para un porcentaje alto de créditos fiscales), si los indicadores muestran que de forma constante un bajo rendimiento, debe ser momento de revisar la construcción de los modelos.

No solo un indicador bajo es motivo para revisar si el modelo está funcionando bien, también si durante algún periodo de tiempo (por ejemplo un año) se generó información nueva con la cual el no fue entrenado, es motivo suficiente para revisar su construcción y volverlo a generar incluyendo los nuevos datos para que esté actualizado.

También es importante mencionar que un modelo debe de tener un ciclo de maduración, no debe generarse uno nuevo solo porque dio un resultado inesperado en algún momento, eso puede deberse a un comportamiento atípico de la información que está calificando en ese momento.

Por lo tanto, el decidir en qué momento se debe calibrar el modelo actual o construir uno nuevo debe ser un trabajo conjunto entre quien genera y quien utiliza el modelo. Esto permitirá que los resultados obtenidos sean confiables.

IV.4. Recapitulación.

El modelado dentro un proyecto de minería de datos, en ocasiones es una cuestión hasta cierto punto artesanal, se tienen que estar atento a los detalles y revisar con cuidado los resultados que se presentarán para la solución de los problemas, así como tener mucha experiencia para tener intuición en la generación de los mismos.

En este capítulo vimos como los datos pueden hacer variar totalmente los resultados de los modelos, por ejemplo para devoluciones se tienen modelos ganadores con apenas el 50% de los casos correctos, esto es un modelo que no pronostica bien.

Un modelo que se acerque al 80% de precisión debe ser considerado un muy buen modelo, sin embargo la precisión no lo es todo en cuanto al resultado.

Si por ejemplo el modelo de devoluciones se acotara a un universo de bajo riesgo, montos menores a X por ciento, donde X sean los importes bajos pero que tienen un mayor número de casos, entonces la utilidad del modelo sería la velocidad con la cual se podrían resolver las solicitudes más comunes y que quitan más tiempo de análisis entre los dictaminadores.

Por otro lado, el modelo de cobranza no solo podría predecir con una precisión bastante aceptable si un crédito fiscal se pagará, esto podría ayudar no solo en revisar las características de los créditos pagables (con una atención pronta con pocas etapas de cobranza) sino también estimar un monto general que se espere cobrar de la cartera activa.

No olvidemos que los datos utilizados en el capítulo fueron generados de manera aleatoria y por lo tanto no tienen un fundamento real, eso se reflejó bastante en el modelo de devoluciones con una precisión de apenas el 50%, pero no para el de cobranza con más del 74%.

Los resultados de los modelos pueden ser implementados dentro de la organización en sistemas ya establecidos a los cuales solo deben asegurar que se tengan las variables que se ocupan en cada uno, mediante PMML podrían generar resultados particulares para cada contribuyente o resultados globales mediante cargas en bloque de un conjunto de contribuyentes.

Por último se debe tener en cuenta el desempeño del modelo en producción para saber en qué momento éste debe de ser calibrado, esto es parte del ciclo de vida de los modelos.

Conclusiones.

El objetivo de este trabajo fue mostrar cómo la minería de datos a través de sus algoritmos predictivos puede utilizarse para la toma de decisiones dentro de una organización.

Se presentó una breve introducción a esta disciplina, explicando de manera muy general los diferentes algoritmos tanto descriptivos como predictivos. Se mostraron los procesos del SAT en donde se aplicarían modelos predictivos: la cobranza y las devoluciones, la idea de plantear el trabajo sobre estos procesos, fue porque representaban los ingresos y egresos que se tienen dentro de la institución.

Se utilizó información pública para explicar cómo funcionan los procesos dentro de la organización, esto es importante porque es el primer paso para empezar un proyecto: entender el problema. Posteriormente utilizando información ficticia debido al secreto fiscal, se generaron datos de manera aleatoria para poder empezar con el entendimiento y preparación de los datos.

En el entendimiento de los datos vimos como empezar a tener el primer contacto con la información, generar reportes para tener una idea de cómo construir la estructura que utilizará el modelo predictivo.

Para la preparación de los datos generamos los cruces, las variables existentes y nuevas, la variable dependiente (objetivo) e independientes (de entrada), esto con la finalidad de tener una vista minable, ésta es una tabla que se genera con toda la información necesaria para empezar con la construcción de modelos.

Para el modelado se utilizó la información construida anteriormente, se empezó tipificando los datos, esto con el fin de que los diferentes algoritmos utilicen correctamente el contenido de las variables de entrada con sus características propias. Se utilizaron herramientas que permiten evaluar de manera rápida todos los algoritmos potenciales a utilizar para la variable objetivo.

Se encontraron correlaciones entre variables para descartar redundancia, se mostró alguna técnica de cómo elegir la más adecuada. Se indicó como balancear las muestras para que los modelos puedan entrenarse con datos que no estén desbalanceados. Se generaron las muestras de entrenamiento para construir el modelo y de comprobación para evaluar los resultados.

Se realizó una comparación de los resultados para encontrar los modelos que tuvieron mejor desempeño, en base a esto se obtuvo el modelo ganador. Se revisó como se generan los resultados de los modelos con datos que ya no son parte de la historia sino reales (recordemos que son ficticios para este trabajo), a los cuáles se les aplicó los modelos ganadores y se mostró la salida de datos que generan.

Se mencionó que los modelos pueden ser exportados a formatos (PMML) que se pueden implementar en lenguajes de programación, esto permite generar sistemas particulares o que pueden incorporarse a otros ya existentes. Se mencionó que los modelos generados aún pueden ser mucho más robustos si se les enriquece con mayor información que permita a los algoritmos ser más precisos, se pueden utilizar no solo los predictivos sino los descriptivos que apoyen a los primeros para un mejor desempeño.

Se recalcó que el generar modelos predictivos permite ver hacia el futuro, predecir con una estimación de probabilidad un problema planteado. Los resultados de los modelos generados pueden ser implementados para mejorar tiempos de respuesta en solicitudes de saldos a favor (para devoluciones) y para dar un valor esperado de recuperación para la cartera activa de créditos fiscales (en cobranza).

Los beneficios al contribuyente en el caso de las devoluciones podría ser el tener resoluciones más rápidas a sus solicitudes y dentro de la institución tener tiempo para revisar los saldos a favor más riesgosos. En cuanto a la cobranza, el tener una aproximación del valor real de recuperación de la cartera de créditos fiscales activos permitiría planear la recaudación en base a un importe esperado y no con respecto al importe total.

Mi experiencia de más de 11 años dentro del SAT es que se tiene mucha información almacenada con un gran potencial pero que se usa poco, se es más reactivo que proactivo, esto hace que los procesos dentro de la institución no tengan un mejor desempeño. Puedo mencionar sin dar más detalles (debido a la confidencialidad de la información) que he participado en dos proyectos similares a los aquí planteados, ambos han tenido éxito y buena aceptación en su implementación dentro de las áreas sustantivas.

El continuar con este tipo de modelos permitirá una mejor toma de decisiones en la administración de riesgos en la institución. Esto podrá ayudar a los objetivos que el SAT tiene que cumplir para el buen funcionamiento en la resolución de saldos y recaudación de impuestos del país.

Bibliografía.

1. ADRIAANS, P. and ZANTINGE, D.. (1996). Data Mining. USA: Addison-Wesley.

2. EZ, C.L. y SANTÍN, D.G. . (2006). Data Mining. Soluciones con Enterprise Miner. México: Alfaomega Grupo Editor.

3. HAN, Jaiwei, KAMBER, Micheline and PEI, Jian. Data Mining. (2011). Data Mining: Concepts and Techniques. USA: Morgan Kaufman.

4. HERNÁNDEZ, José, RAMÍREZ, José y FERRI, César. (2004). Introducción a la Minería de Datos. España: Pearson Prentice Hall.

5. PÉREZ, C.L. y SANTÍN, D.G. (2007). Minería de Datos, Técnicas y Herramientas. España: Ediciones Parinfo.

6. SAT. (2008). Guía de Estudio para la Asignatura de Formación e Información Tributaria. México: SAT Civismo Fiscal.

7. SAT. (2011). Compilación de Legislación Fiscal y Aduanera 2012. México: Dofiscal Editores.

8. SAT. (2013). Asesoría Especializada en Adeudos Fiscales . Enero 2014, de SAT Sitio web: http://www2.sat.gob.mx/adeudosfiscales/

9. SAT. (2013). Devoluciones y Compensaciones. Enero 2014, de SAT Sitio web: http://www.sat.gob.mx/informacion_fiscal/devoluciones_compensaciones/Paginas/default.aspx