PROPUESTA DE USO DE UN MODELO PREDICTIVO PARA LA DETERMINACIÓN DEL PERFIL DE RIESGO DE LOS CONTRIBUYENTES EN LA REPÚBLICA ARGENTINA

Beatriz Steinberg

RESUMEN

La fiscalización del comportamiento de las personas y las empresas juega un rol central en la ejecución de las políticas que le competen a la AFIP.

Este trabajo es justamente una propuesta de mejora del proceso de fiscalización, centrándose en la optimización de la detección de grupos de riesgo. Plantea agregar a la batería de recursos con que ya cuenta la AFIP, el uso de Redes Neuronales para el establecimiento de perfiles de riesgo, orientando el proceso y volviendo más rigurosa la clasificación de contribuyentes en función a potenciales incumplimientos.

La Autora: *Licenciada en Ciencias de la Computación y Licenciada en Psicología de la Universidad de Buenos Aires; Magíster en Administración Estratégica de Negocios de la Universidad de Palermo; Magíster en Ingeniería del Software con titulación conjunta del Instituto Tecnológico de Buenos Aires y la Universidad Politécnica de Madrid; Analista de Fiscalización del Departamento de Sistemas y Diseño de Datos de la Subdirección de Fiscalización, AFIP; Profesora Asociada en la Universidad de Palermo, en la Facultad de Ingeniería y en la Maestría en Ingeniería de Software.*

CONTENIDO

En la actualidad la mayoría de los países vienen trabajando en la formalización de los procesos de planeamiento estratégico de sus administraciones públicas en general, y de las administraciones tributarias en particular.

De un análisis comparativo publicado por la Administración Federal de Ingresos Públicos (AFIP) en enero del 2007 - que abarca Argentina, Chile, Costa Rica, México, Colombia, Nicaragua, Perú, República Dominicana, Australia, Estados Unidos, Canadá, Venezuela, Irlanda, Holanda, Brasil- surge que las administraciones tributarias consideradas, en su mayoría, elaboran planes estratégicos con la tendencia de orientar la Misión de las organizaciones a brindar un servicio de calidad a los contribuyentes, lograr la aplicación eficiente de las leyes y procurar el cumplimiento voluntario de las obligaciones fiscales.

En concordancia con esa Misión, surge la tendencia de fijar la Visión estratégica en el logro de un servicio de calidad para el contribuyente y en la modernización de las organizaciones basándose principalmente en la utilización de nuevas tecnologías y sistemas de información y un plantel calificado. Para hacer efectiva la Visión planteada, los Objetivos estratégicos se centran en la necesidad de optimizar el control y combatir la evasión, la mejora del servicio brindado al contribuyente, el incremento de la eficiencia en la gestión de la organización, la aplicación de nuevas tecnologías a procesos y sistemas, y el desarrollo de los recursos humanos.

En la ejecución de las políticas que le competen a la AFIP y que surgen de los Objetivos estratégicos antes mencionados, juega un rol central la fiscalización del comportamiento de las personas y las empresas.

El primer eslabón en el proceso de fiscalización lo constituyen las tareas de Investigación, al analizar casos susceptibles de tener un interés fiscal significativo. Los casos investigados que revisten interés son sometidos a tareas de fiscalización que pueden o no culminar con éxito, entendido éste como ajustes practicados y cobrados –se apunta a la conformidad del contribuyente y la consecuente evitación de controversias, tanto en sede administrativa como judicial- y sanciones que puedan ser aplicadas.

Tanto la investigación como las distintas modalidades de fiscalización existentes consumen tiempo y recursos humanos; la elección de iniciar y continuar algunas de ellas implica no poder abordar otras.

En ese contexto se instala la importancia de hacer eficiente el proceso en su totalidad, esto es iniciar y proseguir aquéllas investigaciones que van a culminar exitosamente. Esta problemática no es desconocida por la AFIP, que invierte esfuerzos y recursos en la utilización extendida de herramientas informáticas, tanto en la calificación de los contribuyentes atendiendo a su posible riesgo como en la implementación de procesos de control de gestión.

1. LA PROPUESTA

Este trabajo surge de la detección de una necesidad – la que se les plantea a las administraciones tributarias al momento de evaluar estrategias para la promoción del cumplimiento tributario -, reúne y procesa una serie de datos, y obtiene un modelo que permite definir y calificar a un grupo determinado de contribuyentes – los Grandes Nacionales- según su perfil de riesgo.

La elección del universo de análisis se funda en lo reducido del grupo – alrededor del 0,03 % de los contribuyentes totales - y en el interés que genera en la AFIP por su participación en la recaudación, mayor al 48%.

La propuesta se encara en el marco de una metodología de trabajo apta para este tipo de desarrollos: CRISP-DM 1.0, metodología jerárquica que proporciona una descripción del ciclo de vida de un proyecto de minería de datos.

Las inquietudes sobre la viabilidad y la conveniencia de la propuesta encuentran respuesta en las siguientes afirmaciones:

- Las administraciones tributarias poseen en almacenamientos permanentes un gran volumen de datos
- Existe una alta correlación entre el refuerzo de las tareas de fiscalización y la disminución de la evasión
- La percepción del grado de compor-tamiento fiscal de los demás contri-buyentes y la impunidad de los grandes defraudadores operan como justificación de la evasión.
- La caracterización de los Grandes Contri-buyentes Nacionales como cumplidores pragmáticos - en cada momento, deciden si cumplen o no en función de un cálculo egoísta de oportunidad o al resultado de la ecuación tiempo invertido en cumplir versus beneficio a obtener al cumplir; para ellos una administración más eficiente en su tarea fiscalizadora se convierte en la herramienta más eficaz para mejorar el cumplimiento de las obligaciones fiscales.
- La tendencia a minimizar o eludir la carga tributaria por parte de los contribuyentes ante la ausencia de tareas de control efectivas
- La necesidad de la identificación, lo más temprana posible, de prácticas vinculadas con, cuanto menos, incumplimientos de la normativa vigente.
- La concepción en la AFIP de la fiscalización como un proceso lógico y sistemático que requiere el desarrollo y aplicación de herramientas de procesamiento y análisis de información detallada de sujetos, transacciones y operaciones, para identificar segmentos sobre los cuales aplicar acciones específicas que conduzcan al aumento del cumplimiento voluntario y a la detección y prevención de maniobras delictivas, evasivas y elusivas.

1.1 El carácter innovador de la propuesta

Este trabajo plantea un proceso de innovación -entendida como la aplicación de nuevas ideas a viejos problemas, siempre buscando mejoras significativas en la eficiencia, efectividad y calidad- y la posibilidad del uso de las Tecnologías de la Información y la Comunicación (TIC) para viabilizar ese proceso.

Toda innovación implica riesgos, que pueden minimizarse. En este sentido la innovación asociada a este trabajo tiene las siguientes características, todas ellas reductoras de riesgo:

a. Es un refinamiento de un proceso ya existente.
b. No extiende los cambios a nuevas áreas.
c. No es un cambio radical con respecto a las prácticas actuales.

d. Está formulado con claridad de objetivos y límites precisos.
e. No parece compleja su implementación superada la etapa de investigación.
f. El costo de aplicación es inexistente y se esperan beneficios importantes.

1.2 Las tecnologías de la información y la comunicación (TIC´s)

Se está asistiendo al paso de una sociedad analógica –con desplazamiento de objetos reales, como papel- a una digital, en las que se desplazan bits a través de redes de banda ancha y en la que la información en formato texto, voz e imagen se unifica en el concepto de multimedia. Según Juan Hernández "Hablar de las Tecnologías de la Información y las Comunicaciones (TIC) al Servicio de las Administraciones Tributarias es equivalente a hablar de las TIC al servicio de la facilitación y eficiencia en la ejecución de los procesos en las organizaciones".

En función de ello no es llamativo que entre los temas técnicos tratados en todas las Conferencias Técnicas del CIAT, desde 1997 a la fecha, aparezcan trabajos relacionados con las TIC.

Con respecto a la vinculación entre las TIC y las administraciones tributarias englobadas en el CIAT, en la mayoría de ella hoy ya no se discute el uso de la información en la lucha contra la evasión tributaria. Todas ellas trabajan con distintas fuentes de información, ya sea las prestadas por los propios contribuyentes -por medio de declaraciones exigibles por ley- o las obtenidas de terceros -por medio de convenios, cumplimiento de regímenes informativos, acuerdos internacionales o sistemas integrados de información-. Todas se plantean como requisitos esenciales de la información la calidad y la seguridad. Todas priorizan el uso de Internet. La mayor parte de ellas cuenta con herramientas de TI específicas desarrolladas para combatir la evasión.

Al igual que en los otros ámbitos, la introducción de las TIC en las administraciones tributarias requiere colocarlas al servicio de los objetivos; en este caso de lo que se trata es de identificar la manera de volver más eficientes los procesos para la gestión de los impuestos; para lograrlo se debe hacer coincidir el marco estratégico y operativo en la que se desarrollan las Tecnologías de Información con los objetivos estratégicos de la AFIP. Esa coincidencia requiere políticas de tecnología de información explícitas, ya que su ausencia permite el establecimiento de políticas implícitas que, por lo general, resultan potencialmente nocivas para la organización, puesto que sus fundamentos no están claros en todos los casos, no están documentadas (o lo están en forma muy precaria) y generalmente responden a los intereses de los proveedores de tecnología de información, que pretenden crear mercado cautivo.

Una vez que las acciones han sido tomadas se impone la medición de su incidencia en la organización, o sea detectar si han impactado en los objetivos estratégicos de la organización o si se reducen a una mera mecanización, con o sin reducción de costos. En las Administraciones Tributarias hay criterios agregados para la citada medición:

a. La consideración por parte de la sociedad en general y los contribuyentes en particular, sobre la eficiencia, transparencia y credibilidad de las administraciones.
b. El aumento del cumplimiento voluntario por parte de los contribuyentes.
c. El aumento de la recaudación.
d. La necesidad de cumplir con la exigencia de publicación de servicios accesibles por la sociedad.
e. La construcción de un reservorio útil, a partir del gran volumen de datos que se posee, para sistemas predictivos.

Por ello un aprovechamiento intensivo de las TIC no debe renunciar a su uso para la expansión de la capacidad analítica de las administraciones

tributarias, transformando datos e información en materia prima de procesos de búsqueda de conocimiento, de forma de ayudar a mejorar el desarrollo de políticas y la toma de decisiones. En esa línea aparece la Minería de Datos.

1.3 La Minería de datos

La Minería de Datos puede definirse como la exploración y el análisis, por medios automáticos o semiautomáticos de datos para descubrir patrones y reglas; la descripción precedente, al utilizar el concepto de "descubrimiento" apunta a que los patrones y reglas deben estar ocultos hasta ese momento, no ser conocidos y que no es necesario contar con preguntas o intuiciones previas para llegar a ellos. En el mismo sentido Jiawei Han remarca las características que deben tener los patrones y reglas: ser no triviales, previamente desconocidas, implícitas en los datos y potencialmente útiles. El énfasis en la idea de descubrimiento obliga a repensar el rol de la verificación como formando parte de la taxonomía de la minería de datos.

Las administraciones tributarias se hallan entre los más grandes productores, recolectores, consumidores y difusores de información de cada país. El poseer grandes cantidades de datos almacenados en forma persistente, las coloca en condiciones de apelar a procedimientos automáticos o semiautomáticos para encontrar en esos datos conocimiento oculto hasta el momento e interesante: patrones ocultos, asociaciones, cambios, anomalías y estructuras significativas en los datos. La gran capacidad computacional que poseen, sumada al citado volumen de datos, las habilita para encarar procesos de Minería da Datos.

La Minería de Datos, por lo ya dicho, es entonces un proceso posterior a la obtención de los datos, que busca generar información similar a la que podría producir un experto humano, que resulte útil y comprensible; es un eslabón en un proceso más amplio de producción de conocimiento y consiste en la aplicación de algoritmos para la extracción de patrones, utilizando para ello los datos previamente disponibles, que adquieren así más valor.

Dado que se va a utilizar Redes Neuronales parece conveniente alguna aproximación al tema. Se trata de la adaptación de modelos de interconexión de neuronas en el cerebro a la computación digital; las redes neuronales quedan definidas por su topología (organización y disposición de las neuronas de la red en capas), mecanismo de aprendizaje (creación y destrucción de conexiones entre neuronas, así como variación de sus pesos tratando de minimizar el error), el tipo de asociación entre la información de entrada y de salida (hacía adelante, hacía atrás, recurrentes, cualquier combinación de ellas) y la forma de representación de los datos y las salidas (valores continuos, discretos). Se utilizan para problemas de clasificación, estimación y detección de patrones.

1.4 Segmentación

El planteo del problema formulado alude a Grandes Contribuyentes Nacionales. Ello supone una segmentación previa, el agrupamiento de los elementos del universo en estudio en segmentos homogéneos con respecto a criterios previamente definidos, que son justamente los determinantes de la segmentación.

El concepto de segmentación en el caso de las administraciones tributarias, generalmente, apunta a identificar, en base a un concepto de confiabilidad -definido desde la óptica tributaria, aduanera y de la seguridad social-, aquellos segmentos sobre los cuales ejecutar acciones de control diferenciadas, oportunas, razonables y económicas. Se trata de contribuyentes para los que se definen procedimientos especiales, tanto de atención como de fiscalización.

En los países miembros del CIAT este proceso de segmentación ofrece dos niveles: mientras la totalidad de ellos ofrece la partición "informática" de los contribuyentes para ofrecerles un servicio diferenciado con arreglo a su tamaño, actividad, régimen de tributación, naturaleza de

las principales rentas por ellos obtenidas, etc., algunos llevan el concepto de segmentación hacia la organización por tipos de contribuyentes, como sucede en la Argentina.

La necesidad de segmentación surge en la AFIP ante la detección de un pequeño grupo de contribuyentes, con características bien diferenciadas, y una altísima participación en la recaudación. Las características diferenciadas de este grupo son las siguientes:

- Complejidad de las operatorias con impacto impositivo.
- Propensión a litigar (cuentan con asesoramiento de profesionales pertenecientes a estudios jurídicos y/o contables de gran envergadura).
- Rechazo de los ajustes detectados en etapa de inspección: implica iniciar el procedimiento de determinación de oficio en la mayoría de las inspecciones.
- No aceptación de la mayoría de las resoluciones dictadas y apelación ante el Tribunal Fiscal de la Nación.

La respuesta de la AFIP es una segmentación que se manifiesta a varios niveles.

El correlato estructural de la segmentación es la creación de la Subzona Central, incorporada a la Estructura Orgánica de la Dirección General Impositiva por Decreto N° 1.745/74, que luego da lugar a la Dirección Grandes Contribuyentes Nacionales por la Resolución 278/87.

A nivel Informático, la segmentación adquiere entidad con el Sistema DOS MIL - Sistema de Control Diferenciado Especial-, que se constituyó en un intento de minimizar la evasión y el incumplimiento fiscal de los contribuyentes de mayor interés fiscal y que descentraliza la captura de información en los lugares en que se produce. A fines del año 2006 es absorbido por el sistema 2000 regional. Finalmente a partir de julio del 2008 el sistema denominado cuentas tributarias -destinado a registrar y brindar información relativa a deudas y créditos de los contribuyentes y responsables, así como los medios utilizados para su cancelación- es obligatorio para los Grandes Contribuyentes Nacionales mientras el 2000 regional permanece sólo para la administración de obligaciones previas a esa fecha.

A nivel de análisis de riesgo, por RG 1974/2005, modificada por RG 2166/2006, se aprueba el sistema informático “Sistema de Perfil de Riesgo (SIPER), a efectos de categorizar a los contribuyentes y/o responsables -previamente divididos en grupos por volumen de operaciones y actividad económica- de acuerdo con el grado de cumplimiento de sus obligaciones fiscales formales y/o materiales, en cinco (5) categorías o segmentos (A, B, C, D y E), en orden creciente indicativas del riesgo de ser fiscalizado (Categoría A: bajo riesgo de ser fiscalizado; categoría E: alto riesgo de ser fiscalizado). Y justamente este sistema es el punto de partida para la propuesta de este trabajo.

2. DESARROLLO

2.1 Bases para el desarrollo

Es conveniente explicitar los factores críticos de éxito, tanto de la solución propuesta desde la perspectiva del negocio como los del proceso de Minería de Datos.

Son factores críticos de éxito de la solución propuesta la maximización de la recaudación - que debe traducirse en una baja del incumplimiento en el segmento de Grandes Contribuyentes Nacionales-, la mejora de la imagen externa de la AFIP -medible según

el número de contribuyentes que aceptan / aprecian el desempeño y el número de casos en que se cuestiona el perfil de riesgo asignado-, la prevención de fraude - que debe dar lugar a un creciente número de auditorías exitosas sugeridas por la herramienta y un aumento del monto cobrado- y los gastos involucrados en el proyecto -medible según la relación entre recursos empleados e ingresos tributarios conseguidos-.

En cuanto a factores críticos de éxito del proceso propiamente dicho, lo son las medidas de eficiencia propias de los modelos, la aceptación por parte de los expertos de dominio de los resultados y el despliegue de los resultados hacia la comunidad.

Las herramientas a utilizar son las disponibles de escritorio y, para el descubrimiento de patrones, Weka 3.6.1 (Acrónimo de Waikato Environment for Knowledge Analysis, producido por la University of Waikato, New Zealand). WEKA es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario. Está constituido por paquetes de código abierto -integrables a cualquier proyecto con posibilidad de ser enriquecidos con nuevos algoritmos por los usuarios- que incluyen tanto técnicas iniciales de preprocesado de los datos, como de clasificación, agrupamiento, asociación, y, finalmente, visualización de los resultados.

2.2 Comprensión de los datos

La recolección de los datos iniciales se ve notablemente simplificada dado el alto nivel de informatización de la AFIP, que registra en su base de datos centralizada todas las novedades de los contribuyentes y unificadas, toda vez que ello es posible, alrededor de la Clave Única de Identificación Tributaria. Los datos iniciales a explorar son los que conforman archivos de trabajo trimestrales para atender los requerimientos del Sistema de determinación de Perfil de Riesgo, que reúnen, por cada contribuyente y en un único registro, todos los datos que hacen a su comportamiento tributario.

Los datos así recolectados son 3547 registros, con 88 campos por registro, conteniendo datos que hacen al comportamiento tributario de los Grandes Contribuyentes Nacionales durante los tres cuatrimestres del año 2009.

En cuanto a las características de las variables

- La mayoría son de tipo categórico e indican presencia o ausencia de desvío.
- Las hay numéricas discretas (cantidad de causas penales, cantidad de empleados) y algunas continuas (deuda) que, por la gran cantidad de valores que ofrecen deberán ser tratadas al momento de ser utilizadas.
- Se detecta la existencia de valores fuera de rango en determinadas variables, que deberán ser tratadas como discretas, construyendo rangos y agrupando en uno de ellos a todos los pocos valores excesivamente grandes.
- Hay valores faltantes en todos los casos para determinados atributos, que representan situaciones que los responsables han decidido no seguir relevando, por lo que se los decide eliminar
- La preponderancia de ciertos valores en ciertos atributos y en todas las categorías, que lleva a suponer que la variable resultará poco predictiva
- Se detecta atributos redundantes, que se decide eliminar
- Se detecta la presencia de algunos datos que no se aplican a todo el universo en estudio; en ese caso el cálculo de correlaciones con respecto a la clase circunscribiéndose al grupo a que se aplican arroja coeficientes de correlación no muy distintos a los obtenidos trabajando el universo total de Grandes Nacionales, por lo que no se considera el tema un problema.

En cuanto a la semántica de los datos

- Los contribuyentes del universo seleccionado están conformados por un 36,48% de personas físicas y un 63,52% de personas jurídicas, de las que la mayoría, un 89,3%, son sociedades anónimas.
- En un primer acercamiento a los datos es posible detectar que el porcentajes de contribuyentes sancionados o con incumplimientos detectados es bajo y menor al porcentaje que resulta de considerar al universo total de contribuyentes. Y que la cantidad de juicios contenciosos en trámite (30%) es alta frente a los finalizados a favor de la AFIP, aún en forma parcial. (5% y 2% respectivamente).

2.3 Preparación de los datos

Definida la variable de clase, el uso preliminar de la herramienta weka aporta luz sobre los atributos más significativos usando una serie de evaluadores de selección, lo que sumando al uso de correlaciones permite eliminar atributos con muy bajo valor predictivo.

Los valores faltantes para un atributo no requieren la construcción de valores especiales; por el contrario, la inexistencia de valores, cuando se da, lejos de ser un problema referido a un valor desconocido, es pertinente y marca un hecho real (por ejemplo, la ausencia de presentación de declaraciones juradas en un no obligado).

El volumen de registros con que se cuenta no hace necesario ni aconsejable trabajar con muestras.

Se crea una serie de nuevos atributos que agrupan y ponderan los desvíos construyendo índices.

Se construyen juegos de datos diferenciales en los que el atributo a predecir es numérico o categórico y en el que los a tributos independientes son nominales, usando S/N cuando hay dos opciones y Aceptable/Regular/Malo cuando se trabaja con tres opciones para asignar puntaje al desvío.

2.4 Modelado preliminar

En forma preliminar se intenta crear un esquema de agrupamiento de los contribuyentes incluidos en este estudio mediante el análisis de cluster, con la convicción de que aumentará el conocimiento de los datos disponibles y que ello es un buen punto de partida para toda búsqueda posterior de patrones ocultos en los datos.

Este intento de encontrar una agrupación natural entre las instancias consideradas de acuerdo a la similitud que presenten entre ellas las variables observadas tiene implícita la expectativa de que el agrupamiento que se busca, al trabajar sobre atributos que indican desvíos en el comportamiento fiscal, resulte solidario con la actual clasificación según perfil de riesgo y deje a los “buenos contribuyentes” en algún grupo, así como a los “regulares” y “malos” en otros.

Ello no resulta así pese al uso de la herramienta bajo distintas modalidades.

Se puede suponer que la asignación de categorías por el usuario, al definir cortes fijos y posiblemente arbitrarios en determinados puntajes, deja de distinto lado de la “frontera” a instancias demasiado similares, que un algoritmo de agrupamiento deja en el mismo grupo. A si mismo la idea de trabajar por puntaje, también sumariza a aquellos desvíos con muy baja incidencia, que una rutina de agrupamiento no considera. De las experiencias realizadas se concluye que el análisis de cluster no permite generar agrupamientos solidarios con las categorías hoy definidas por el usuario.

2.5 Redes neuronales

El intento en este caso es utilizar Redes Neuronales para la clasificación de los contribuyentes en las cinco categorías definidas previamente por la AFIP. Para ello, a partir de un la existencia de un número finito de clases y

su asignación a un conjunto de datos de entrenamiento, se trata de construir un modelo para cada clase que pueda ser utilizado para la clasificación de datos futuros

La parametrización elegida, el método de testeo durante la construcción del modelo y el comando usado son los siguientes:

decay	FALSE	Provoca el decrecimiento de la tasa de aprendizaje original, ayudando a evitar divergencias
autoBuild	TRUE	Agrega y conecta los niveles ocultos de la red.
hiddenLayers	-H a	Define la cantidad de nodos de los niveles ocultos de la red, separados por comas. Admite comodines 'a' = (atributos + clases) / 2, 'i' = atributos, 'o' = clases , 't' = atributo + clases
learningRate	-L 0.3	Tasa de aprendizaje o proporción en que los pesos son modificados
Momentum	-M 0.2	Momentum aplicado a los pesos durante la modificación
nominalToBinary Filter	<-B/~~b~~> <T/F>	Preprocesa las instancias con un filtro. Aumenta la performance si hay atributos nominales en los datos. Resulta irrelevante en este caso
normalizeAttributes	TRUE	Normaliza los atributos, aun los nominales, entre -1 y 1, para aumentar la performance
normalizeNumeric Class	<-C/~~b~~> <T/F>	Normaliza la clase, si es numérica, y sólo en forma interna, entre -1 y 1. Resulta irrelevante en este caso
reset	TRUE	Permite que el proceso recomience automáticamente con una menor tasa de aprendizaje si se detecta divergencia.
Seed	-S 0	Usado para inicializar el generador de números random para el seteo de los pesos iniciales de las conexiones entre nodos.
trainingTime	-N 500	Número de ciclos para el entrenamiento.
validationSetSize	-V 0	Porcentaje del set de validacion set Si no es 0, el entrenamiento continúa hasta que el error en el set de validación se reduce o los ciclos de entrenamiento se cubren. Si es 0, no se usa set de validación y el entrenamiento dura el numero de ciclos indicado
validationThreshold	-E 20	Usado para terminar la validación. El valor indica cuantos tiempos en una instancia el error debe reducirse para que el entrenamiento termine.
Test Options	Cross Validation 10 Folds	
java weka.classifiers.functions.MultilayerPerceptron -t totalweka.arff -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a -G -R -d modeloclasifica.out		

El modelo resultante, con 21 nodos, posee un bajo nivel de error y un alto nivel de cubrimiento, según se muestra en la **Cuadro 1,** que también discrimina ambos conceptos por cada clase.

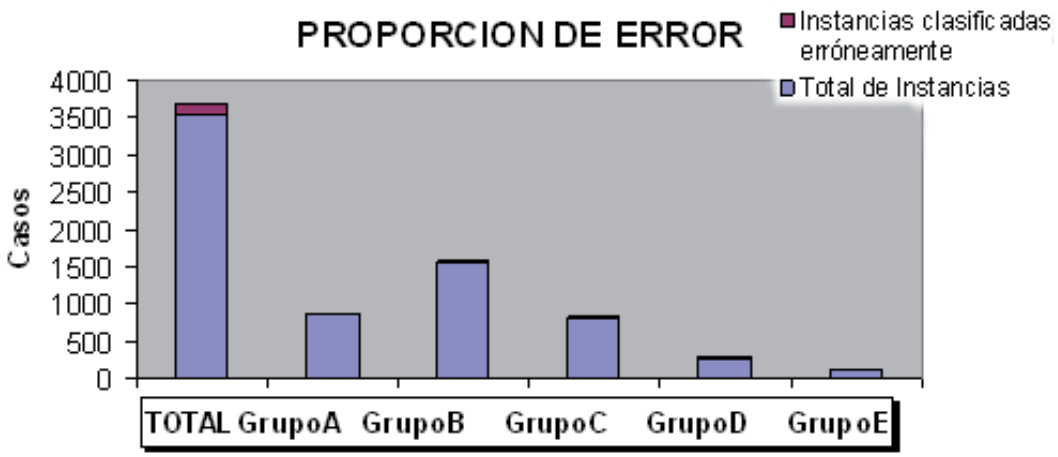
Cuadro 1

Proporción de error en el Modelo

CASOS	CUBRIMIENTO		PRECISIÓN	
A: 857	0.989	848/857	0.977	20/868
B: 1543	0.985	1520/1543	0.975	39/1559
C: 799	0.939	750/799	0.955	39/789
D: 251	0.849	213/251	0.869	32/245
E: 97	0.804	78/97	0.907	8/ 86
Total : 3547	**0.97**		0.**961**	

Gráfico 1

Proporción de error en el Modelo



3. EVALUACIÓN

La evaluación es una etapa insoslayable en un proyecto de Minería de Datos y está sujeta a una serie de condiciones, determinadas por el tipo de modelo (descriptivo o predictivo), el negocio al que el modelo se aplica, los objetivos iniciales y la intención del destinatario del modelo. No sólo atiende a cuestiones técnicas, sino del negocio y puede poner en descubierto cuestiones del tipo de detección de patrones no importantes para el negocio, pobreza en cuanto al conocimiento generado, "sobre aprendizaje" o necesidad de enriquecer los datos básicos en cuanto a tamaño de registros o atributos.

En este caso, dado que se trata de lograr clasificar correctamente a un universo de contribuyentes en función de su cumplimiento, la precisión es fundamental; por ello la métrica elegida para poner a prueba el comportamiento del modelo es la determinación del porcentaje de tuplas mal clasificadas y se aplica a un nuevo juego de datos, 1362 registros relativos al primer trimestre del año 2010, sobre el que se realiza la misma preparación que la comentada anteriormente.

Es también importante destacar que al momento de realizar la evaluación se debe establecer si todos los errores de clasificación tienen igual peso o si es más grave evaluar como sin riesgo en cuanto a cumplimiento fiscal a un contribuyente realmente riesgoso que considerar como de alto riesgo a un contribuyente que no lo es.

El resultado de la evaluación es de casi 70 % de aciertos (916 casos sobre 1362)

La primera cuestión a analizar para entender los errores del modelo es comparar la incidencia de las categorías propuestas por el modelo con la incidencia real de esas categorías.

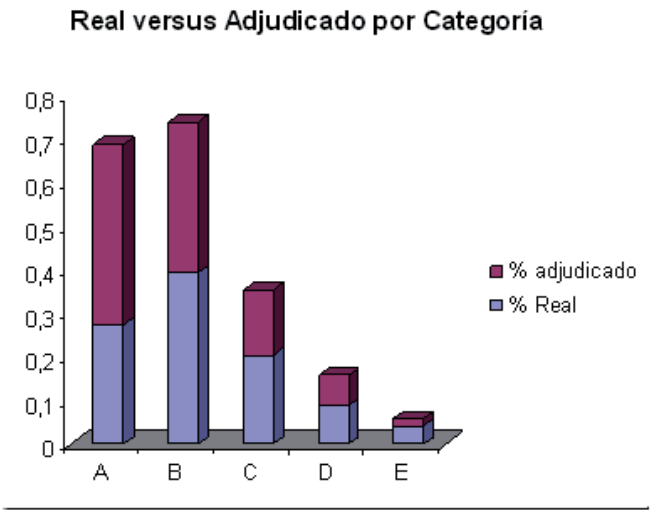
Cuadro 2

Incidencia de cada categoría luego de la aplicación del modelo

CATEGORIA	INCIDENCIA REAL	ADJUDICACION
A	0,27312775	0,41409692
B	0,3928047	0,34801762
C	0,20484581	0,15051395
D	0,08810573	0,06975037
E	0,04111601	0,01762115

Gráfico 2

Incidencia real y adjudicada de cada categoría en la evaluación



El modelo parece ser "generoso" al momento de determinar el nivel potencial de incumplimiento de los contribuyentes, lo que puede explicarse por la presencia de desvíos con muy poco soporte, que el algoritmo de clasificación no considera. La herramienta al ubicar a los contribuyentes en una categoría mejor a la que le asignan los usuarios, incrementa la participación de la categoría A en el universo, llevándola de un 27% a un 41 % y decrementa la participación de la categoría E, llevándola de un 4% a un 2%.

Se impone luego evaluar para cada categoría, los porcentajes en que el modelo acierta y aquellos en los que se equivoca, discriminando si el error es en el sentido de dotar de menor riesgo al contribuyente (califica mejor) de dotarlo de mayor riesgo (califica peor).

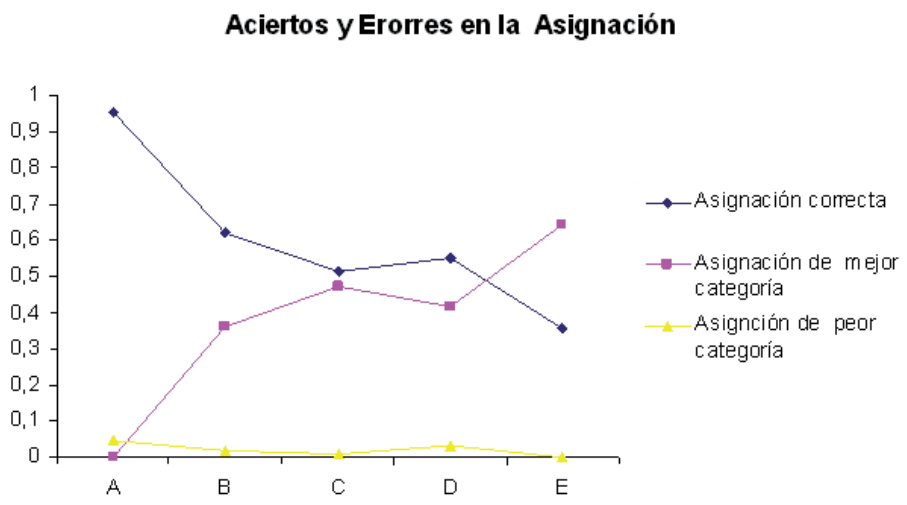
Cuadro 3

Detalle de la categorización de instancias en la Evaluación

CATEGORIA	ACIERTA	CALIFICA MEJOR	CALIFICA PEOR
A	0,9516129	0	0,0483871
B	0,61869159	0,36074766	0,02056075
C	0,51612903	0,47311828	0,01075269
D	0,55	0,41666667	0,03333333
E	0,35714286	0,64285714	0

Gráfico 3

Aciertos y errores en la evaluación



Nuevamente se visualiza que los errores tienen más que ver con la ubicación en una categoría mejor a la asignada por los usuarios, que con una peor ubicación; en este último caso el error es menos significativo ya que solo ocurre en un 4% de casos de la categoría A, un 2% de casos la Categoría B, un 1 % de casos la categoría C y un 3 % de casos de la Categoría D; en total, este error representa menos del 3% considerado sobre el universo utilizado en la evaluación del modelo.

Se hace evidente que el modelo mejoraría su precisión si se generasen casos ficticios con presencia de los desvíos de escaso soporte.

4. CONCLUSIONES

La optimización de las tareas de fiscalización en las administraciones tributarias es esencial; si bien el objetivo de mediano plazo es aumentar el cumplimiento voluntario de las obligaciones, en el corto plazo es esencial aumentar el nivel de cumplimiento; en ese contexto la fiscalización es una pieza fundamental, y su optimización una necesidad.

Las administraciones tributarias están abocadas a su logro, que es posible y viene de la mano de un cambio de cultura que no busca aumentar el número de inspecciones o de acciones de fiscalización, sino orientarlas cualitativamente en el sentido de obtener mejores resultados con menor consumo de recursos; se asienta concretamente en una serie de pilares, entre los que se puede mencionar la segmentación de contribuyentes, la aplicación de medidas diferenciales a esos segmentos, la aparición de áreas especializadas para su atención, las políticas de detección temprana de fraude, los análisis del contexto político y económico en que los contribuyentes se desenvuelven y del impacto de la globalización de la economía, los aportes de la sociología fiscal, la construcción de perfiles de riesgo utilizando en forma centralizada e integrada toda la información de que se dispone, los cambios organizacionales.

El control de riesgo se ubica en el centro de la nueva orientación de las tareas de fiscalización; más que la gestión a posteriori se apunta a la optimización de la detección de grupos de riesgo.

Surge así la necesidad de contar con herramientas automáticas que puedan hacer su aporte en el establecimiento de perfiles de riesgo de los distintos grupos. Para viabilizar innovaciones el uso de las TIC es fundamental, ya que las mismas no podrían pensarse, dados los volúmenes de información a procesar y las áreas geográficas a cubrir, sin su uso. En esa línea se inscribe la creciente presencia de las TIC en las administraciones de los distintos países incluidos en el CIAT y la reorientación de su uso, que pasó de ser un simple auxiliar de cálculo a convertirse en un facilitador del cambio cultural, que las ubica en la mecánica de comunicación dominante en la sociedad, les facilita los intercambios con otros organismos nacionales e internacionales y les permite cumplir con la facilitación de las tareas a los contribuyentes.

En ese contexto, este trabajo se centra en la construcción de modelos para la descripción y clasificación según el riesgo de incumplimiento de los Grandes Contribuyentes Nacionales –el grupo de más interés para la AFIP-, así como de la búsqueda de reglas que permitan explicar tanto el mantenimiento como la variación de la clasificación previa.

Particularmente, se logra aplicar redes neuronales sobre los datos existentes en la AFIP relativos a los desvíos registrados por los Grandes Contribuyentes Nacionales durante tres períodos del año 2009.

Por otro lado, se procede a definir un ambiente experimental para validar los resultados, con el objetivo de evaluar la efectividad y el éxito de la solución propuesta. Para ello, se utilizan como medida de rendimiento el grado de precisión, que se mide como el porcentaje de tuplas mal clasificadas.

Las pruebas realizadas usando el modelo propuesto permiten demostrar que es posible aplicar algoritmos de clasificación y contar con un modelo de predicción de riesgo sobre los contribuyentes de mayor interés para las administraciones tributarias; el grado de confianza encontrado en este trabajo resulta del 70% y es superior al obtenido con otras herramientas de minería de datos; por caso la obtención de reglas mediante árboles que clasifiquen a futuros contribuyentes en las 5 categorías en función de sus desvíos arroja sólo un 61% de coincidencias.

A futuro, la obtención de datos con mayor cobertura temporal puede permitir el uso de series temporales en la búsqueda de patrones secuenciales; el completamiento de la fuente de datos genuina con la creación de casos de prueba que cubran todo el universo de los desvíos puede aumentar el nivel de precisión de los modelos predictivos; y trabajar sobre asociación entre distintos desvíos puede orientar la investigación, ante la aparición de algunos de ellos, de la presencia de otros que se dan junto a los primeros.

En síntesis,

Es posible optimizar el proceso de fiscalización utilizando criterios innovativos, sin mayores riesgos, con la colaboración de TIC, dotando a las administraciones tributarias de modelos útiles para la determinación de perfiles de riesgo.

La utilidad para las administraciones públicas de contar con esos tipos de modelos es múltiple: hacen un uso efectivo y útil de los grandes volúmenes da datos que poseen, posibilitan acceso universal a la información centralizada, garantizan transparencia, calidad y seguridad e igual tratamiento ante igual conducta y mejora la imagen que de ellas poseen los contribuyentes.

5. BIBLIOGRAFÍA

AFIP, Dirección de Planeamiento y Análisis de Gestión, Dto. Planeamiento: Proceso de Formulación del Plan Estratégico, (Buenos Aires, 2007).

ANAO, Australian National Audit Office: Innovation in the Public Sector - Enabling Better Performance, Driving New Directions Better Practice Guide, (Australia, 2009).

BERRY, M. y LINOFF, G.: Data Mining techniques for Marketing, Sales and Customer Support, (USA, 1977).

BRITOS, Paola Verónica y GARCÍA MARTÍNEZ, Ramón y HOSSIAN, Alejandro y SIERRA, Enrique: Minería de Datos basada en Sistemas Inteligentes, (Buenos Aires, 2005).

CHAPMAN, P. y CLINTON, J y KERBER, R y KHABASA, T. y REINARTZ, T y SHEARE, C. y WIRH, R: CRISP-DM 1.0 Step-by-step data mining guide in CRISP-DM consortium, (EEUU, 2000).

CONDE, Alberto: Formulación de una política informática para la administración tributaria, en Una Administración Tributaria para el Nuevo Milenio – Escenarios y Estrategias - Conferencias Técnicas del CIAT, (Washington, 2000).

ESPER, Susana C: Factores subjetivos incidentes en la institucionalización de la e-taxation, en Cuadernos del Instituto AFIP, Vol. nº 8, (Buenos Aires, 2009).

ESTÉVEZ, Alejandro: La administración tributaria frente al cambio tecnológico, en Cuadernos del Instituto AFIP, Vol. nº 8, (Buenos Aires, 2009).

HAN, J. y KAMBER, M.: Data mining: Concepts and techniques, (EEUU, 2006).

HERNÁNDEZ BATISTA, Juan: Las tecnologías de la información y de las comunicaciones al servicio de las administraciones tributarias, en Una visión moderna de la administración tributaria - 43ª Asamblea General del CIAT, (Santo Domingo, 2009).

MAIMON, Oded Y ROKACH, Lior: Data Mining and Knowledge Discovery Handbook, (EE.UU, 2005).

MICHALSKI, R.S. y BRATKO, I. y KUBAT, M.: Machine Learning and Data Mining. Methods and Applications, (EE.UU., 1998).

RUSSO, Marcos: Administración de riesgos en Aduana-Una perspectiva Horizontal, en Cuadernos del Instituto AFIP, Vol. nº 10, (Buenos Aires, 2010).

SEGARRA TORMO, Santiago: "El uso de las nuevas tecnologías para facilitar el cumplimiento fiscal ", en La administración tributaria al servicio del ciudadano - Conferencias Técnicas del CIAT, (Sevilla, 2001).

STEINBERG, Beatriz: Aplicación de Minería de Datos al proceso de fiscalización de Grandes Contribuyentes Nacionales en la Administración Federal de Ingresos Públicos, en La Tributación en las Sociedades Digitales Nativas - XXI Encuentro Técnico Internacional de Administradores Fiscales en Argentina, (Mar del Plata, 2011).

Todo el material de esta publicación fue preparado, e impreso en la Secretaría Ejecutiva del CIAT, Apartado 0834-02129, Panamá Rep. de Panamá. Se terminó la impresión en el mes de Junio de 2012.