

Minería de datos aplicada a la acción social en la provincia de Salta

G. R. Rivadera¹

Resumen

Actualmente es más difícil para el gobierno provincial poder satisfacer en forma oportuna y eficaz la demanda por ayuda social de amplios sectores de la población. Una de las mayores limitaciones a este esfuerzo es la falta de repositorios de información sobre las necesidades de las personas, y por consiguiente, de modelos de análisis adecuados que ayuden a la identificación de necesidades no evidentes en la población carente de recursos.

Este trabajo de investigación propone una solución desde la informática, implementando un sistema que recoja información relevante de las actividades de acción social en el municipio de Rosario de Lerma, las integre de forma adecuada en un almacén de datos (*Datawarehouse*) y las analice con modelos adecuados de minería de datos, de forma de ayudar en la planificación estratégica, organización y control de las actividades de acción social que se realizan desde el municipio. Sobre todo se centra en los modelos de análisis de minería de datos, aporte novedoso en este ámbito, ya que no existen trabajos similares en nuestro país. En este informe, y por cuestiones de espacio, nos centraremos básicamente en el estudio de minería de datos.

Palabras Clave : almacenes de datos – acción social – análisis de datos – minería de datos

Introducción

Actualmente es cada vez más difícil para el gobierno provincial poder satisfacer en forma oportuna y eficaz la demanda por ayuda social de amplios sectores de la población. Una de las mayores limitaciones a este esfuerzo es la falta

de repositorios de información sobre las necesidades de las personas, y por consiguiente, de modelos de análisis adecuados que ayuden a la identificación de necesidades no evidentes en la población carente de recursos. Este proyecto de investigación propone el diseño de un sistema de gestión y análisis de datos de

¹ Facultad de Ingeniería e Informática, UCASAL,

ayuda social en la provincia de Salta. Este sistema engloba por una parte diseñar e implementar un subsistema de registro y gestión de los datos, con un paradigma orientado a objetos, utilizando la metodología Métrica, Versión III, implementado en el lenguaje C# de Microsoft. Por otra parte, se pretende el diseño de un subsistema de análisis multidimensional de los datos, y un subsistema de análisis mediante técnicas de minería de datos.

Estado de la cuestión

Estado de los sistemas de gestión social en Argentina y en particular en la Provincia de Salta

Actualmente la Republica Argentina tiene índices de pobreza que superan el 25%. Este problema se traduce en que muchos de sus habitantes no tienen los suficientes recursos para alcanzar un nivel de vida adecuado a los estándares mundiales. Si bien se entiende mucho del progreso social y económico del país se lograría con adecuadas fuentes de trabajo e ingreso económico, surge la necesidad de paliar o ayudar a las personas a corto plazo mediante programas de acción social. Los gobiernos han implementado, durante muchos años, tanto a nivel nacional como provincial, muchos programas de este tipo. Estos programas intentan paliar la situación de pobreza de las personas en forma de ayudas alimentarias y materiales. Sin

embargo, a lo largo de la historia, siempre existieron varios problemas en la realización de los mismos. Podemos sintetizarlos en los siguientes párrafos:

- Alcance: Un problema grave es que no todos los programas alcanzan a la población que realmente los necesita. Muchas veces, por cuestiones políticas y de partidismo, se otorgan ayudas a personas que son cercanas al grupo de poder de turno, o en cambio se dan a las personas que mas frecuentan los centros de distribución, o las que reinviden con más insistencia.
- Oportunidad: Otro problema es el hecho de que, de llegar la ayuda social a la persona que la necesita, en algunos casos lo entregado no corresponde con la necesidad real de la persona. Es decir, por dar un caso, si la persona convive en un núcleo habitacional saturado, con muchas otras personas, al carecer de una planificación e información adecuada, se le brinda la ayuda menos urgente, por no decir equivocada, y se le entrega por ejemplo una caja de comida, u otras cuestiones ajenas a su problemática esencial de ese momento, que es la habitacional.
- Falta de pro-actividad: La mayoría de los programas sociales son reactivos. Esto es, por lo general se informa a la población de que se está ayudando de tal o cual forma, y se la cita en un lugar determinado lugar, donde se hace entrega de la ayuda social en sus varias formas. Sin embargo, muchas veces es necesario implementar programas proactivos, esto es, que resuelvan los

problemas sociales in-situ, y que se acerquen a los que más lo necesitan, que como se dijo en el inciso a) no son los que necesariamente están en el centro de distribución de la ayuda social.

De alguna forma, este trabajo pretende aportar un principio de solución a los problemas mencionados, desde la óptica de la ingeniería del software y el análisis de datos, dado que entendemos que la falta de información oportuna y el análisis de los mismos ocasionan estos problemas.

Modelos de minería de datos estudiados

Introducción

Este proyecto de investigación estaba estructurado en tres partes bien diferenciadas. La primera consistía en implementar un sistema transaccional de carga de datos de encuestas y asistencias sociales. La segunda implicaba la creación de un repositorio de datos o *Datawarehouse*, a partir del cual se podría analizar los datos mediante técnicas dimensionales. La tercera, parte, tal vez la más importante del proyecto, era la creación de modelos de minería de datos, tomados a partir de vistas minables del *datawarehouse* creado en la etapa anterior.

Todas las etapas del proyecto fueron realizadas con éxito. En particular en la última etapa, se desarrollaron las siguientes tareas de minería de datos, en base a la metodología CRISP-DM elegida para el desarrollo de esta etapa²:

- Definición de los objetivos de estudio para la etapa de minería de datos del proyecto
- Exploración estadística y general de los datos del almacén de datos
- Limpieza de los datos
- Análisis de agrupamiento (*clustering*) de la población
- Análisis de agrupamiento (*clustering*) sobre datos de asistencias a la población
- Planificación estratégica de las ayudas del área de Acción Social en lo referido a las proyecciones necesarias sobre el otorgamiento de los diferentes tipos de ayudas a las personas con *Microsoft Time Series* (MTS)
- Exploración de datos de asistencias usando el algoritmo *Naïve Bayes*
- Exploración de datos de asistencias usando el algoritmo de clasificación y predicción de árboles de decisión (Clasificación con *Microsoft Decision Trees*)
- Estudio de la secuencia temporal de peticiones de asistencias de las personas (algoritmo *Microsoft Sequence Clustering*)

² Por cuestiones de espacio también es imposible realizar una explicación detallada de CRISP-DM. Para mayor información sobre la misma se puede recurrir a

- Exploración de las relaciones entre los distintos tipos de peticiones de ayudas de las personas, análisis de reglas de asociación (*Association Rules*)
- Análisis estadístico exploratorio de datos de satisfacción de atención del personal
- Análisis de agrupamiento (*clustering*) sobre datos de satisfacción de atención del personal
- Métodos para la evaluación de los modelos presentados

Podemos clasificar los estudios realizados también de esta forma, según el tipo de análisis realizado, y las técnicas utilizadas:

·Análisis de agrupamiento (*Clustering*)

- Sobre datos de asistencias a la población
- Sobre datos de satisfacción de atención del personal del área
Análisis sobre datos de asistencias
- Planificación estratégica de las ayudas del área de Acción Social en lo referido a las proyecciones necesarias sobre el otorgamiento de los diferentes tipos de ayudas a las personas con *Microsoft Time Series* (MTS)
- Exploración de datos de asistencias usando el algoritmo *Naïve Bayes*
- Exploración de datos de asistencias usando el algoritmo de clasificación y predicción de árboles de decisión (Clasificación con *Microsoft Decision Trees*)

- Estudio de la secuencia temporal de peticiones de asistencias de las personas (algoritmo *Microsoft Sequence Clustering*)
- Exploración de las relaciones entre los distintos tipos de peticiones de ayudas de las personas, análisis de reglas de asociación (*Association Rules*)

A Continuación explicaremos sólo uno de los estudios realizados, el de agrupamiento de la población con redes neuronales SOFM, ya que por cuestiones de espacio, sería imposible describir todos los estudios realizados en este artículo. De todas formas, es posible consultar los demás estudios en los informes de avance presentados y en la memoria final del proyecto de investigación, en el Consejo de Investigaciones de la UCASAL.

Análisis de agrupamiento (*Clustering*)

Sobre datos de la población (demografía)

Para este análisis se usaron las técnicas de Redes neuronales con mapas auto organizados (SOM) y agrupamiento mediante *K-Means*. Los datos se analizaron con la herramienta *Statistica*, versión 8, un software propietario comercial del que se dispone una versión de prueba.

En la fig. 1 y 2 se muestran las dos pantallas donde se especifica el tamaño

de la red, los ciclos de entrenamiento, y otros parámetros relacionados con la inicialización de la red. Inicialmente se había definido una red de 9x9, sin embargo, se terminó usando una red de 11x11 ya que en las diferentes pruebas

realizadas con anterioridad a la que se muestra, se encontró que esta es la configuración con menos error, como muestra la Fig 3. En la Fig. 4 se muestra el resultado de la ejecución del algoritmoSOFM con la red de 11x11.

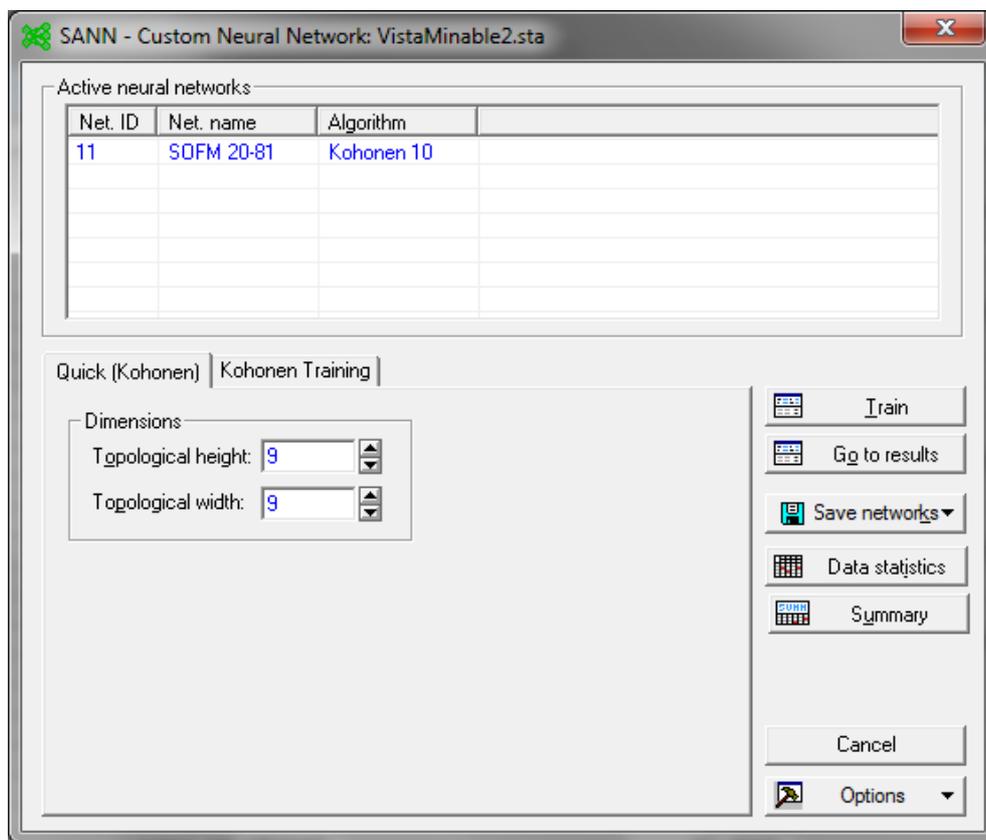


Fig.1 Ejemplo de parametrización de Kohonen en Statistica (ejemplo 9x9)

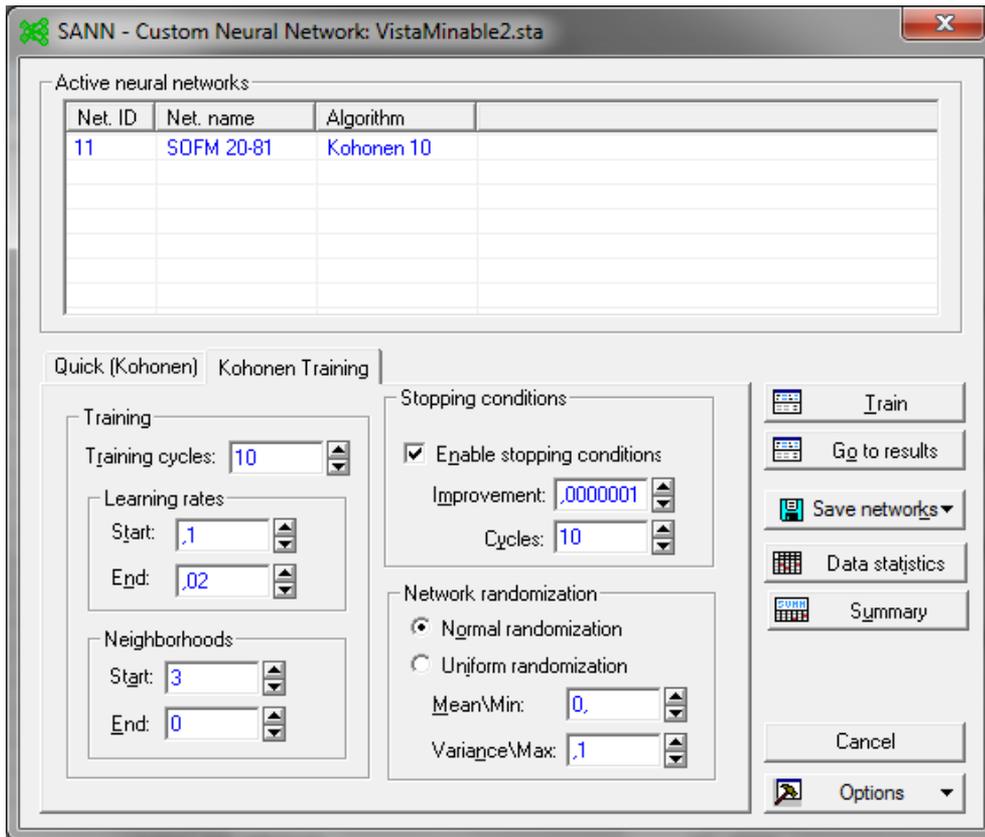


Fig.2 Parámetros adicionales en la parametrización de Kohonen (ejemplo 9x9).

Summary of active networks (VistaMinable2.sta)				
Index	Net.name	Training error	Test error	Training algorithm
1	SOFM 194-81	3,043309	2,913447	Kohonen10
Summary of active networks (VistaMinable2.sta)				
Index	Net.name	Training error	Test error	Training algorithm
2	SOFM 194-100	2,923745	2,759579	Kohonen10
Summary of active networks (VistaMinable2.sta)				
Index	Net.name	Training error	Test error	Training algorithm
3	SOFM 194-121	2,874457	2,732012	Kohonen10

Fig. 3 Error encontrado en varias aproximaciones.

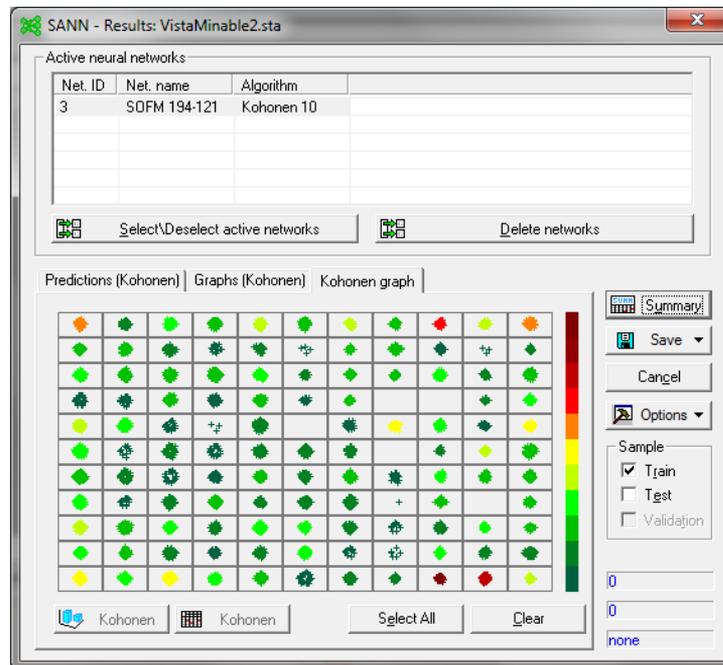


Fig.4 Resultado del algoritmo en forma gráfica.

		Weights spreadsheet (VistaMinibaja2.sta) Network: 12.SOFM 20-8-1								
Inputs	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)	(1, 7)	(1, 8)	(1, 9)	
ZonaBarrio	0,361861	0,342403	0,357666	0,700808	0,053163	1,000000	0,000000	1,000000	0,9997	
Sexo	0,587156	0,395612	0,742165	0,320087	1,000000	0,992506	0,000000	0,985428	0,9999	
EstadoCivil	0,802173	0,485472	0,848366	0,355366	1,000000	0,960592	0,000000	0,995324	0,1579	
Edad	0,561395	0,517390	0,660448	0,625157	1,000000	0,217690	0,000000	0,001457	0,1798	
Parentesco	0,250010	0,377672	0,840515	0,321239	1,000000	0,021141	0,000000	0,000000	0,1505	
CobrasigUnivesal	0,503059	0,363926	0,079057	0,159424	1,000000	0,000000	0,000000	0,424454	0,1585	
CobralinterTabaco	0,413174	0,355647	0,098933	0,429661	1,000000	0,000000	0,000000	0,990938	0,3106	
Discapacidad	0,408513	0,058926	0,177160	0,450109	1,000000	0,000000	0,103021	1,000000	0,2465	
Enfermedad	0,360747	0,071146	0,336685	0,724005	0,544235	0,437183	0,309683	1,000000	0,1429	
CoberturaSalud	0,419181	0,107497	0,599246	0,698297	0,000000	0,341024	0,505175	0,999867	0,1428	
Desnutrido	0,379564	0,108534	0,705099	0,372074	1,000000	0,218965	0,669685	0,996835	0,1447	
Ocupacion	0,648087	0,498287	0,710550	0,186701	1,000000	0,022742	0,565590	0,067825	0,1440	
Trabajaja	0,602049	0,444419	0,585779	0,533789	1,000000	0,024862	0,483091	0,767472	0,1480	
NoOcupacion	0,731419	0,636272	0,569413	0,367789	0,993713	0,000000	0,291529	0,000000	0,1655	

Fig. 5 Ejemplo de tabla de pesos resultantes en la red

3.2.2 Evaluación de los resultados:

Al realizar el análisis sobre los datos obtenidos de la red de Kohonen, nos encontramos con la gran cantidad de neuronas a analizar, así se decidió exportar la matriz de clasificación a la base de datos original de SQL Server, y desde allí diseñar el procedimiento almacenado que se muestra a continuación,

para obtener una idea general de los datos relevantes en cada agrupamiento. La idea fue utilizar la moda, como el valor que más se repite en cada grupo para las variables categóricas, y otros indicadores estadísticos para las continuas, y además analizar algunos grupos de especial interés. La tabla obtenida del análisis con este procedimiento se muestra más abajo (Tabla 1).

```
CREATE PROCEDURE AnalizarKohonen AS
BEGIN
CREATE TABLE #Aux(
[Neuron ID] int,
Cantidad int,
EdadPromedio float,
EdadMaxima float,
EdadMinima float,
MedianaEdad float,
IngresoMensualProm float,
IngresoMensualMax float,
IngresoMensualMin float,
MedianaIngreso float,
[Neuron location] nvarchar(50),
ZonaBarrio nvarchar(100),
Sexo nvarchar(100),
EstadoCivil nvarchar(100),
Parentesco nvarchar(100),
```

. Rivadera

```
Oficio    nvarchar(100),  
BolsonAlimentario nvarchar(100),  
Ocupacion nvarchar(100),  
Trabaja   nvarchar(100),  
NoOcupacion nvarchar(100),  
FuenteIngresos nvarchar(100),  
AporteJub nvarchar(100),  
CobraAsigFam nvarchar(100),  
CobraAsigUniversal nvarchar(100),  
CobraInterTabaco nvarchar(100),  
Discapacidad nvarchar(100),  
Enfermedad nvarchar(100),  
CoberturaSalud nvarchar(100),  
Desnutrido nvarchar(100))
```

```
DECLARE Cur CURSOR
```

```
READ_ONLY
```

```
FOR
```

```
SELECT
```

```
DISTINCT [Neuron ID], [Neuron location]
```

```
/* [Activation], Edad, IngresoMensual, [Neuron  
location], ZonaBarrio, Sexo,
```

```
EstadoCivil, Parentesco, Oficio, BolsonAlimentario, Ocupacion, Trabaja, No
```

```
Ocupacion, FuenteIngresos,
```

```
AporteJub, CobraAsigFam, CobraAsigUniversal, CobraInterTabaco, Discapacidad,
```

```
Enfermedad, CoberturaSalud, Desnutrido*
```

Minería de datos aplicada a la acción social en la provincia de Salta

```
WHILE (@@fetch_status <> -1)
BEGIN
    IF (@@fetch_status <> -2)
    BEGIN
        INSERT INTO #Aux SELECT @NeuronID, NULL AS Cantidad, NULL
AS EdadPromedio, NULL AS EdadMaxima, NULL AS EdadMinima, NULL AS
MedianaEdad,
        NULL AS IngresoMensualProm, NULL AS IngresoMensualMax, NULL
AS IngresoMensualMin, NULL AS MedianaIngreso, @Neuronlocation AS
Neuronlocation,
        NULL AS ZonaBarrio, NULL AS Sexo,
        NULL AS EstadoCivil, NULL AS Parentesco, NULL AS Oficio, NULL
AS BolsonAlimentario, NULL AS Ocupacion, NULL AS Trabaja, NULL
AS NoOcupacion,
        NULL AS FuenteIngresos, NULL AS AporteJub, NULL AS
CobraAsigFam, NULL AS CobraAsigUniversal, NULL AS
CobraInterTabaco, NULL AS Discapacidad,
        NULL AS Enfermedad, NULL AS CoberturaSalud, NULL AS
Desnutrido
        --Mediana ZonaBarrio
        UPDATE #Aux SET ZonaBarrio = (
        SELECT TOP 1 WITH ties ZonaBarrio
        FROM Kohonen11
        WHERE ZonaBarrio IS Not NULL AND [Neuron ID] = @NeuronID
        GROUP BY ZonaBarrio
        ORDER BY COUNT(*) DESC)
        WHERE [Neuron ID] = @NeuronID
        --Mediana Sexo
        UPDATE #Aux SET Sexo = (
```

. Rivadera

```
SELECT TOP 1 WITH ties Sexo
FROM Kohonen11
WHERE Sexo IS Not NULL AND [Neuron ID] = @NeuronID
GROUP BY Sexo
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID
```

-EstadoCivil

```
UPDATE #Aux SET EstadoCivil = (
SELECT TOP 1 WITH ties EstadoCivil
FROM Kohonen11
WHERE EstadoCivil IS Not NULL AND [Neuron ID] = @NeuronID
GROUP BY EstadoCivil
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID
```

-Parentesco

```
UPDATE #Aux SET Parentesco = (
SELECT TOP 1 WITH ties Parentesco
FROM Kohonen11
WHERE Parentesco IS Not NULL AND [Neuron ID] = @NeuronID
GROUP BY Parentesco
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID
```

-Oficio

```
UPDATE #Aux SET Oficio = (
```

Minería de datos aplicada a la acción social en la provincia de Salta

```
SELECT TOP 1 WITH ties Oficio
FROM Kohonen11
WHERE Oficio IS Not NULL AND [Neuron ID] = @NeuronID
GROUP BY Oficio
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID

--BolsonAlimentario
UPDATE #Aux SET BolsonAlimentario = (
SELECT TOP 1 WITH ties BolsonAlimentario
FROM Kohonen11
WHERE BolsonAlimentario IS Not NULL AND [Neuron ID] =
@NeuronID
GROUP BY BolsonAlimentario
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID

--Ocupacion
UPDATE #Aux SET Ocupacion = (
SELECT TOP 1 WITH ties Ocupacion
FROM Kohonen11
WHERE Ocupacion IS Not NULL AND [Neuron ID] = @NeuronID
GROUP BY Ocupacion
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID

--Trabaja
UPDATE #Aux SET Trabaja= (
SELECT TOP 1 WITH ties Trabaja
```

. Rivadera

```
FROM Kohonen11
WHERE Trabaja IS Not NULL AND [Neuron ID] = @NeuronID
GROUP BY Trabaja
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID
```

-NoOcupacion

```
UPDATE #Aux SET NoOcupacion = (
SELECT TOP 1 WITH ties NoOcupacion
FROM Kohonen11
WHERE NoOcupacion IS Not NULL AND [Neuron ID] = @NeuronID
GROUP BY NoOcupacion
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID
```

-FuenteIngresos

```
UPDATE #Aux SET FuenteIngresos = (
SELECT TOP 1 WITH ties FuenteIngresos
FROM Kohonen11
WHERE FuenteIngresos IS Not NULL AND [Neuron ID] =
@NeuronID
GROUP BY FuenteIngresos
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID
```

-AporteJub

```
UPDATE #Aux SET AporteJub = (
SELECT TOP 1 WITH ties AporteJub
```

Minería de datos aplicada a la acción social en la provincia de Salta

```
FROM Kohonen11
WHERE AporteJub IS Not NULL AND [Neuron ID] = @NeuronID
GROUP BY AporteJub
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID

--CobraAsigFam
UPDATE #Aux SET CobraAsigFam= (
SELECT TOP 1 WITH ties CobraAsigFam
FROM Kohonen11
WHERE CobraAsigFam IS Not NULL AND [Neuron ID] =
@NeuronID
GROUP BY CobraAsigFam
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID

--CobraAsigUniversal
UPDATE #Aux SET CobraAsigUniversal= (
SELECT TOP 1 WITH ties CobraAsigUniversal
FROM Kohonen11
WHERE CobraAsigUniversal IS Not NULL AND [Neuron ID] =
@NeuronID
GROUP BY CobraAsigUniversal
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID

--CobraInterTabaco
UPDATE #Aux SET CobraInterTabaco= (
```

. Rivadera

```
SELECT TOP 1 WITH ties CobraInterTabaco
FROM Kohonen11
WHERE CobraInterTabaco IS Not NULL AND [Neuron ID] =
@NeuronID
GROUP BY CobraInterTabaco
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID

--Discapacidad
UPDATE #Aux SET Discapacidad= (
SELECT TOP 1 WITH ties Discapacidad
FROM Kohonen11
WHERE Discapacidad IS Not NULL AND [Neuron ID] =
@NeuronID
GROUP BY Discapacidad
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID

--Enfermedad
UPDATE #Aux SET Enfermedad= (
SELECT TOP 1 WITH ties Enfermedad
FROM Kohonen11
WHERE Enfermedad IS Not NULL AND [Neuron ID] = @NeuronID
GROUP BY Enfermedad
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID

--CoberturaSalud
UPDATE #Aux SET CoberturaSalud= (
```

Minería de datos aplicada a la acción social en la provincia de Salta

```
SELECT TOP 1 WITH ties CoberturaSalud
FROM Kohonen11
WHERE CoberturaSalud IS Not NULL AND [Neuron ID] =
@NeuronID
GROUP BY CoberturaSalud
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID

-Desnutrido
UPDATE #Aux SET Desnutrido= (
SELECT TOP 1 WITH ties Desnutrido
FROM Kohonen11
WHERE Desnutrido IS Not NULL AND [Neuron ID] = @NeuronID
GROUP BY Desnutrido
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID

-Mediciones de Edad
UPDATE #Aux SET EdadPromedio = (
SELECT CONVERT(INT,AVG(Edad)) FROM Kohonen11 WHERE Edad IS
Not NULL AND [Neuron ID] = @NeuronID),
EdadMaxima=(
SELECT CONVERT(INT,MAX(Edad)) FROM Kohonen11 WHERE Edad IS
Not NULL AND [Neuron ID] = @NeuronID),
EdadMinima =(
SELECT CONVERT(INT,MIN(Edad)) FROM Kohonen11 WHERE Edad IS
Not NULL AND [Neuron ID] = @NeuronID),
Cantidad = (SELECT COUNT(Edad) FROM Kohonen11 WHERE Edad IS
Not NULL AND [Neuron ID] = @NeuronID)
```

. Rivadera

```
WHERE [Neuron ID] = @NeuronID
```

```
UPDATE #Aux SET MedianaEdad = (  
SELECT TOP 1 WITH ties Edad  
FROM Kohonen11  
WHERE Edad IS Not NULL AND [Neuron ID] = @NeuronID  
GROUP BY Edad  
ORDER BY COUNT(*) DESC)  
WHERE [Neuron ID] = @NeuronID
```

-Ingreso Mensual

```
UPDATE #Aux SET IngresoMensualProm = (  
SELECT CONVERT(DECIMAL(18,2),AVG(IngresoMensual)) FROM  
Kohonen11 WHERE IngresoMensual IS Not NULL AND [Neuron ID] =  
@NeuronID),  
IngresoMensualMax= (  
SELECT CONVERT(DECIMAL(18,2),MAX(IngresoMensual)) FROM  
Kohonen11 WHERE IngresoMensual IS Not NULL AND [Neuron ID] =  
@NeuronID),  
IngresoMensualMin=(  
SELECT CONVERT(DECIMAL(18,2),MIN(IngresoMensual)) FROM  
Kohonen11 WHERE IngresoMensual IS Not NULL AND [Neuron ID] =  
@NeuronID)  
WHERE [Neuron ID] = @NeuronID
```

```
UPDATE #Aux SET MedianaIngreso = (  
SELECT TOP 1 WITH ties IngresoMensual
```

Minería de datos aplicada a la acción social en la provincia de Salta

```
FROM Kohonen11
WHERE IngresoMensual IS Not NULL AND [Neuron ID] =
@NeuronID
GROUP BY IngresoMensual
ORDER BY COUNT(*) DESC)
WHERE [Neuron ID] = @NeuronID

END

FETCH NEXT FROM Cur INTO @NeuronID,@Neuronlocation
/
*,@Activation,@Edad,@IngresoMensual,@Neuronlocation,@ZonaBarrio,@Sexo,
@EstadoCivil,@Parentesco,@Oficio,@BolsonAlimentario,
@Ocupacion,@Trabaja,@NoOcupacion,@FuenteIngresos,@AporteJub,
@CobraAsigFam,@CobraAsigUniversal,@CobraInterTabaco,@Discapacidad,
@Enfermedad,@CoberturaSalud,@Desnutrido */

END

CLOSE Cur

DEALLOCATE Cur

SELECT * FROM #Aux order by [Neuron ID]

END
```

3.3.3 Conclusiones de *Clustering* con Kohonen

Se pueden observar varios grupos claramente diferenciados, vinculando los grupos de las diferentes neuronas:

Grupo 1: Mujeres casadas amas de casa sin cobertura social en las zonas 1, 7 y 4.

Grupo 2: Mujeres en concubinato amas de casa sin cobertura social en las zona 1.

Grupo 3: Mujeres solteras, mayormente estudiantes, en las zonas 1, 7 y 4.

Grupo 4: Mujeres casadas amas de casa sin cobertura social en las zonas 4, 5 y 3

Grupo 5: Mujeres solteras sin cobertura social en las zona 1

Grupo 6: Mujeres solteras jefes de hogar sin trabajo ni cobertura en la zona 1.

Grupo 7: Hombres casados con trabajo sin cobertura de salud

Grupo 8: Hombres solteros estudiantes en su mayoría sin cobertura de salud en zonas 1,4,5,6,7.

Grupo 9: Hombre casados jefes de hogar, que no trabajan o jubilados en zona 4.

Grupo 10: Hombre solteros jefe de hogar que tienen trabajo con cobertura en zonas 5,4 y 2.

Grupo 11: Hombres casados jubilados con cobertura de salud PAMI.

Existen muchos otros grupos puntuales, para un análisis exhaustivo se puede analizar sencillamente la matriz presentada anteriormente. De todas formas, esta matriz y el gráfico nos permiten concluir que los grupos de más riesgos serían el 1 y el 6, y que la zona 1 es la que posee una mayor incidencia de ca-

sos de riesgo, y por lo tanto a la que habría que apuntar inicialmente en la cobertura de acción social.

Ver Tabla I

3.3.4 Validación del modelo con K-means

Para validar y contrastar los datos obtenidos anteriormente, usamos el algoritmo de K-medias (*K-means*) del mismo software *Statistica*. Para ello usamos la opción de Generalized Cluster Analysis con K-means, como se muestra en la Fig. 7. A continuación se parametriza el algoritmo, se le coloca una cantidad de Clusters entre 10 y 50 (para que se encuentre en el rango obtenido con SOFM), y además se marca la opción de que realice una validación cruzada (*V-fold cross validation*), Fig. 8.

3.3.5 Análisis y resultados con K-means

En este caso, los resultados se obtienen de forma más directa, ya que el mismo algoritmo nos proporciona la matriz de grupos, que se muestra en la Tabla 2. Se observan grupos similares a los encontrados por la red neuronal, técnica utilizada previamente. En particular se vuelven a observar los grupos de riesgo, de mujeres sin cobertura social en la zona 1.

Ver Tabla II

Zona/ Barrio	Sexo	Estado Civil	Parentesco	A.U ¹	I.T ²	No Ocupación	Fuente Ingresos	AJ ³	AF ⁴	Zona/ Barrio	Sexo	Estado Civil	Parentesco	A.U	I.T	No ocupación	Fuente Ingresos	AJ	A.F	Zona/ Barrio	Sexo
	F	Casado	Cónyuge (casado civilmente)	No	No	Ama de casa	Sin respuesta	No	No	1	F	Casado	Cónyuge (casado civilmente)	No	No	Ama de casa	Sin respuesta	No	No	1	F
1	F	Soltero	Hijo/a	No	No	Estudiante	Sin respuesta	No	No	1	F	Soltero	Hijo/a	No	No	Estudiante	Sin respuesta	No	No	1	F
1	F	Soltero	Hijo/a	No	No	Estudiante	Sin respuesta	No	No	1	F	Soltero	Hijo/a	No	No	Estudiante	Sin respuesta	No	No	1	F
1	F	Soltero	Hijo/a	No	No	Estudiante	Sin respuesta	No	No	1	F	Soltero	Hijo/a	No	No	Estudiante	Sin respuesta	No	No	1	F
1	M	Soltero	Jefe	No	No	Sin datos	Sueldos/ jornales	No	No	1	M	Soltero	Jefe	No	No	Sin datos	Sueldos/ jornales	No	No	1	M
4	F	Casado	Jefe	No	No	Jubilado o pensionado	Jubilado/ pensión ordinaria	No	No	4	F	Casado	Jefe	No	No	Jubilado o pensionado	Jubilado pensión ordinaria	No	No	4	F

Tabla 1. Tabla de grupos (parcial) obtenida a partir de la aplicación del procedimiento almacenado de la Fig. 6.

Referencias

- ¹ Asignación universal
- ² Inter tabaco
- ³ Aportes jubilatorios
- ⁴ Asignación familiar

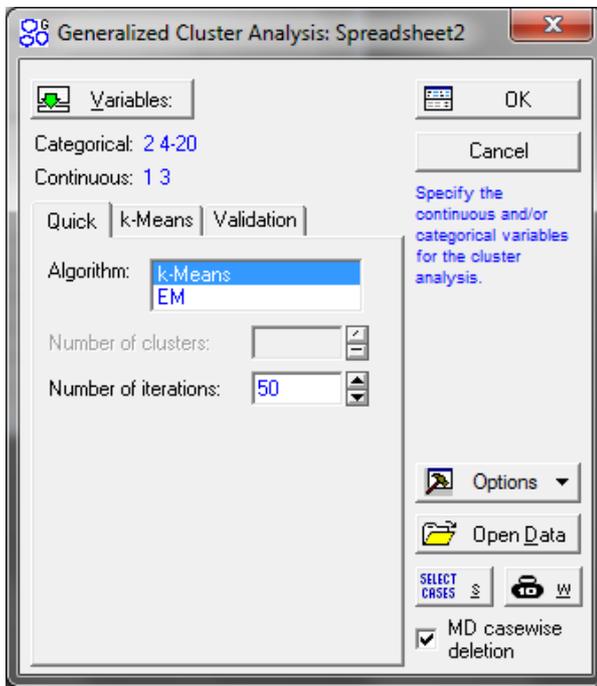


Fig. 7 Selección de la técnica de K-means

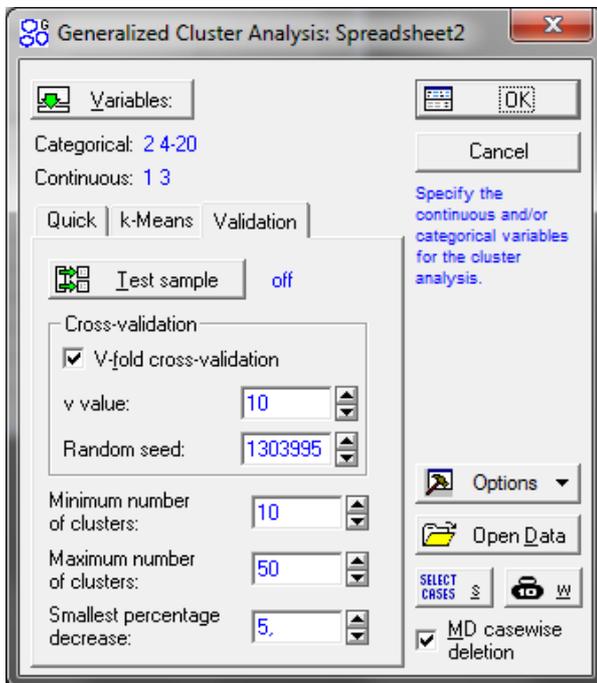


Fig. 8 Parametrización de K-means

Neuron Id	Cant.	Promedio Edad	Edad. Máxima	Edad Mínima	Mediana Edad	Ingreso mensual promedio	Ingreso mensual máximo	Ingreso mensual mínimo	Mediana Ingreso	Neuro Location	Zona/ Barrio	Sexo	Estado civil	Parentesco
1	273	43	69	2	29	2,34	240	0	0	(1,1)	1	F	Casado	Cónyuge casado civilmente
2	60	45	60	27	52	45,75	1200	0	0	(1,2)	7	F	Casado	Cónyuge (casado civilmente)
3	145	42	69	2	36	9,79	335	0	0	(1,3)	4	F	Casado	Cónyuge (casado civilmente)
4	113	32	63	1	29	37,61	1500	0	0	(1,4)	1	F	Unido de hecho	Cónyuge (casado civilmente)

...

...





...

Oficio	Bolsón alimentario	Ocupación	Trabaja	No ocupación	Fuente de Ingresos	Aporte jubilatorio	Asignación familiar	Asignación Universal	Intertabaco	Discapacidad	Enfermedad	Cobertura Social	Desnutrido
No tiene oficio	Sin datos	Sin datos	No	Ama de casa	Sin respuesta	No	No	No	No	Sin discapacidad	Sin enfermedad	Sin cobertura	No
Manualidades y artesanías	Sin datos	Sin datos	No	Ama de casa	Sin respuesta	No	No	No	No	Sin discapacidad	Sin enfermedad	Sin cobertura	No
Sin respuestas	Sin datos	Sin datos	No	Ama de casa	Sin respuesta	No	No	No	No	Sin discapacidad	Sin enfermedad	Sin cobertura	No
Sin respuestas	Sin datos	Sin datos	No	Ama de casa	Sin respuesta	No	No	No	No	Sin discapacidad	Sin enfermedad	Sin cobertura	No

Tabla 2: grupos resultantes con k-means.

2. Conclusiones

Si bien es cierto que cada estudio en particular tiene sus propias conclusiones, se pueden enunciar las siguientes conclusiones generales:

- El estudio de minería de datos sirvió para ayudar a mejorar la distribución de las ayudas sociales.
- El análisis reveló algunas zonas de la ciudad de Rosario de Lerma donde se necesitan más tipos de ayudas de una categoría que de otra.
- El estudio sirvió para realizar una segmentación de la población, útil en las campañas de ayudas sociales.
- El sistema quedó funcionando para registro y análisis de datos.

Referencias

Datawarehousing:

1. Imhoff & Galembo, *Mastering Data Warehouse Design: Relational and Dimensional Techniques*, Wiley Publishing, 2003
2. Inmon, *Building the Data Warehouse*, (Third Edition). John Wiley & Sons, 2002
3. Kimball & Caserta, *The Data Warehouse ETL Toolkit*, Indianapolis, Wiley, 2004.
4. Kimball & Ross, *The Data Warehouse*

Toolkit: The Complete Guide to Dimensional Modeling (Second Edition), New York, Wiley, 2002.

5. Kimball & Ross, *The Kimball Group Reader; Relentlessly Practical Tools for Data Warehousing and Business Intelligence*, Indianapolis, Wiley, 2010.
6. Kimball et al., *The Data Warehouse Lifecycle Toolkit*. 2nd Edition. New York, Wiley, 2008
7. Kimball et al., *The Data Warehouse Lifecycle Toolkit*. New York, Wiley, 1998.
6. Mundy & Thornthwaite, *The Microsoft Data Warehouse Toolkit—With SQL Server 2005 and the Microsoft Business Intelligence Toolset*, Indianapolis, Wiley, 2006.

Datamining

1. Boris Mirkin, *Clustering for Data Mining: A Data Recovery Approach*, Chapman & Hall – CRC, 2005.
2. Rui Xu, Donald C. Wunsch III, *Clustering*, IEEE & Wiley Publishing, 2009.
3. Lewicki Pavel, Hill Thomas, *Statistics, Methods and Applications*, Statsoft, 2008.
4. Paolo Giudici, Silvia Figini, *Applied Data Mining for Business and Industry*, John Wiley & Sons, 2009.
5. Jamie McLennan, ZhaoHui Tang, Bogdan Crivat, *Data Mining With Microsoft SQL 2008*, Wiley Publishing Inc., 2009.