



ARTÍCULOS

Análisis del desempleo urbano a través de un estudio comparativo de métodos de clasificación

Margarita Díaz , Fernando Ferrero , Cecilia Díaz , Patricia Caro y María Inés Stimolo

Revista de Economía y Estadística, Cuarta Época, Vol. 43, No. 2 (2005), pp. 61-85.

<http://revistas.unc.edu.ar/index.php/REyE/article/view/3818>



La Revista de Economía y Estadística, se edita desde el año 1939. Es una publicación semestral del Instituto de Economía y Finanzas (IEF), Facultad de Ciencias Económicas, Universidad Nacional de Córdoba, Av. Valparaíso s/n, Ciudad Universitaria. X5000HRV, Córdoba, Argentina.
Teléfono: 00 - 54 - 351 - 4437300 interno 253.
Contacto: rev_eco_estad@eco.unc.edu.ar
Dirección web <http://revistas.unc.edu.ar/index.php/REyE/index>

Cómo citar este documento:

Díaz, M., Ferrero F., Díaz C., Caro P. y Stimolo M. (2005). Análisis del desempleo urbano a través de un estudio comparativo de métodos de clasificación. *Revista de Economía y Estadística*, Cuarta Época, Vol. 43, No. 2, pp. 61-85.

Disponible en: <http://revistas.unc.edu.ar/index.php/REyE/article/view/3818>

El Portal de Revistas de la Universidad Nacional de Córdoba es un espacio destinado a la difusión de las investigaciones realizadas por los miembros de la Universidad y a los contenidos académicos y culturales desarrollados en las revistas electrónicas de la Universidad Nacional de Córdoba. Considerando que la Ciencia es un recurso público, es que la Universidad ofrece a toda la comunidad, el acceso libre de su producción científica, académica y cultural.

<http://revistas.unc.edu.ar/index.php/index>



REVISTAS
de la Universidad
Nacional de Córdoba



Universidad
Nacional
de Córdoba



FCE
Facultad de Ciencias
Económicas



1613 - 2013
400
AÑOS



Revista de Economía y Estadística Vol. XLIII(2) - Año 2005
Instituto de Economía y Finanzas - Facultad de Ciencias Económicas,
Universidad Nacional de Córdoba - Argentina

Análisis del desempleo urbano a través de un estudio comparativo de métodos de clasificación

MARGARITA DÍAZ

mdiaz@eco.unc.edu.ar

FERNANDO FERRERO

fferrero@eco.unc.edu.ar

CECILIA DÍAZ

cdiaz@eco.unc.edu.ar

PATRICIA CARO

pacaro@eco.unc.edu.ar

MARÍA INÉS STIMOLO

mstimolo@eco.unc.edu.ar

Instituto de Estadística y Demografía, Facultad de Ciencias Económicas, Universidad Nacional de Córdoba

Resumen

Este trabajo propende a identificar los factores de riesgo que inciden en la precariedad laboral de la Población Económicamente Activa. Se adoptó como plataforma informativa la base de datos de la Encuesta Permanente de Hogares, octubre 2002, relevada en las ciudades de Córdoba, Rosario y en el gran Buenos Aires. El efecto de las variables predictoras sobre la condición de actividad del encuestado se estimó a través de los Análisis de Regresión Logística y Árboles de Decisión. Adicionalmente, y

a los fines de mejorar la performance de la clasificación obtenida, se aplicaron los métodos de Redes Neuronales y Vecino más Cercano.

Palabras Claves: Condición de actividad, regresión logística, vecino más cercano, árboles de decisión, encuesta permanente de hogares

Clasificación JEL: C14, C25, J64

Abstract

This work attempts to identify several risk factors underlying the so called precariousness of the argentinian labour force. To this end it was taken into account the database of the Periodically Household Survey, October 2002, a survey regularly carried out in the cities of Córdoba, Rosario and great Buenos Aires. Final effects of predictive variable over the activity condition of the individual being interviewed were modeled through Logit Regression and Tree Decision Models. Additionally, in order to improving the performance of the estimated classification rules two statistical models were also worked out, say, Neuronal Networks and Nearest Neighbour methods.

Key words: Semiparametric and Nonparametric Methods, Multivariate Statistical Analysis, Unemployment: Models, Duration, Incidence, and Job Search, Household Data.

JEL classification: C14, C25, J64

I. INTRODUCCIÓN

La utilización del instrumental multivariado se ha convertido en una herramienta prácticamente insustituible cuando se analiza la acción de un conjunto de factores interactuantes y sus efectos en una o más variables de respuesta.

En el análisis del mercado laboral el estudio de la oferta de trabajo plantea requerimientos metodológicos que son específicamente inherentes a las fuentes estadísticas disponibles. La información adscripta a las unidades de análisis provienen de las encuestas de hogares y estudios de tipo socio-

económico donde prevalecen datos medidos sobre la base de escalas nominales y ordinales, a excepción de variables tales como edad, antigüedad en la ocupación, ingresos y número de componentes del hogar.

Este trabajo apunta a un estudio comparativo de los métodos de clasificación supervisada aplicables a esta problemática y resultará de especial utilidad para los investigadores de la Economía Laboral, habida cuenta que el desempleo es uno de los problemas más acuciantes de la sociedad argentina. A este respecto no se puede soslayar el hecho que en la década de los 90 la desocupación abierta registró un aumento considerable y elevó la tasa de desempleo del 13% en 1998 al 19,4% en Octubre del 2002.

Focalizando el análisis en las áreas metropolitanas de Córdoba, Buenos Aires y Rosario, se analizaron los datos de la Encuesta Permanente de Hogares (EPH) correspondientes a la onda octubre 2002. A los fines de la modelización se aplicaron métodos de clasificación supervisada especialmente indicados cuando las variables predictoras son categóricas: 1) Métodos Parcialmente Paramétricos: Regresión Logística (en sus versiones Binaria, Ordinal y Multinomial y 2) Métodos No Paramétricos, tales como el del Vecino más Cercano y Árboles de Decisión en su versión CHAID. La aplicación de diferentes técnicas a un mismo problema tiene por objetivo analizar la eficacia relativa de los diversos instrumentos del análisis multivariado, al tiempo que permite evaluar y seleccionar aquellos que conduzcan a una mejor descripción del perfil sociodemográfico de desocupados y sub-ocupados.

Si bien en los estudios microeconómicos -realizados a partir de encuestas de hogares- se utiliza frecuentemente la clasificación binaria Ocupado y Desocupado, el deterioro de las condiciones laborales se tradujo sin embargo en la presencia de una alta proporción de personas que trabajan a tiempo parcial o trabajan a tiempo completo pero recibiendo una remuneración muy baja, razón por la cual en esta investigación se optó por segmentar la población económicamente activa en tres grupos: Ocupado, Subocupado y Desocupado, incluyendo en la segunda categoría a las personas ocupadas que trabajan menos de 35 horas semanales o ganan menos de \$200 mensuales.

II. FACTORES INCLUIDOS EN EL ANÁLISIS

En la selección de variables determinantes de la situación laboral, se han considerado tres dimensiones básicas, a saber: a) datos personales del

encuestado, b) características socio-demográficas del hogar, e c) historia laboral de las personas.

En materia de "Datos Personales", se han considerado *género, edad, estado civil, parentesco con el jefe de hogar y nivel de educación alcanzado*. A este respecto, el género constituye una de las manifestaciones de discriminación laboral que aún perviven en la sociedad argentina ya que se observa una proporción significativamente mayor de mujeres en el estamento desocupados. En cuanto a la edad, caben dos alternativas: 1) la mayor tasa de desocupación está en los jóvenes (hasta 24 años) y 2) existe una composición etaria bimodal en el desempleo, compuesta por dos grupos de riesgo ubicados en los extremos de la edad activa (hasta 24 años y más de 50 años).

Por lo que se refiere al *estado civil*, se clasificó a las personas en dos categorías, según tuviesen o no pareja, esperándose mayor riesgo de precariedad laboral para los segundos. Igual tratamiento se realizó en materia de *parentesco*, dividiéndolos en jefes y no jefes de hogar, toda vez que la jefatura de hogar juega un papel muy importante en el estado ocupacional de las personas. En la literatura consultada se considera que los trabajadores secundarios (no jefes) tienen mayores dificultades para conseguir empleo. Por último, el principal indicador del capital humano es la *educación*, ya que cuanto menor sea el nivel educativo, mayor será la chance de tener un empleo precario o estar desocupado.

En otro orden, para la dimensión "Características del hogar", se han considerado los indicadores *tamaño del hogar y Necesidades Básicas Insatisfechas (condición de NBI)*. Con relación a la primera, se supone que las personas económicamente activas que tienen a su cargo un mayor número de integrantes del hogar, realizarán un esfuerzo mayor que los que no están en igual situación, razón por la cual los primeros exhibirán un menor porcentaje de desocupación. Por otra parte es razonable suponer que las personas que habitan en hogares con NBI tendrán una mayor probabilidad de pertenecer al estrato de desocupados.

Finalmente, en la dimensión "Historia Laboral", se incluyeron dos variables que reflejan experiencia laboral, *categoría ocupacional y rama de actividad*, ambas referidas a su último empleo. Respecto a la primera, se supone que los trabajadores con empleo formal son menos vulnerables que los trabajadores por cuenta propia. Por otro lado, el proceso de apertura y liberalización de la economía, implementado en la década del 90, ha castigado más fuertemente a determinadas ramas productivas, entre las que se cuentan la industria y la construcción.

III. MÉTODOS APLICADOS

Para abordar problemas con predictores categóricos se han desarrollado métodos parcialmente paramétricos, entre los cuales el más popular es la *regresión logística*. Con este enfoque, sólo se modela el cociente de densidades, sin asumir una forma funcional específica, requiriendo sólo linealidad para el logaritmo del cociente de densidades (McLachlan, 1992).

Los métodos *no paramétricos* no establecen ningún supuesto acerca de las densidades condicionales a los grupos. Entre ellos, el más aplicado es el *Vecino más Cercano*, que estima la probabilidad condicional de que un elemento pertenezca a un grupo determinado, a través de la proporción de puntos o casos próximos a dicho elemento que corresponden a ese grupo.

El otro método aplicado, es un método recursivo conocido como *Árboles de Decisión*, que generalmente representa una serie de preguntas binarias consecutivas y es construido por divisiones repetidas de subconjuntos en subconjuntos descendentes.

III.1. Regresión Logística Ordinal y Multinomial

Este modelo supone que el logaritmo del cociente de densidades es lineal. En el caso binario, la variable respuesta Y asume los valores (0, 1) y dicho logaritmo se expresa de la siguiente manera:

$$\ln \left\{ \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \right\} = \beta_0 + \beta' \mathbf{x}, \quad (1)$$

donde $f_j(\mathbf{x})$ representa la función de probabilidad para el grupo j ($j = 0, 1$).

Una observación es asignada al grupo 1, que simbolizamos G_1 , si se cumple que $\beta_0 + \beta' \mathbf{x} \geq \ln \frac{\pi_0}{\pi_1}$, donde π_j ($j = 0, 1$) representa la proporción poblacional de cada grupo, por lo que $\pi_0 + \pi_1 = 1$. Las probabilidades a posteriori de pertenencia a los grupos son¹:

$$P(Y = 1 | \mathbf{x}) = \frac{\pi_1 f_1(\mathbf{x})}{\pi_0 f_0(\mathbf{x}) + \pi_1 f_1(\mathbf{x})} = \frac{e^t}{1 + e^t} \quad (2)$$

¹ Aplicando el Teorema de Bayes y operando.

$$pr(Y = 0 / \mathbf{x}) = \frac{\pi_0 f_0(\mathbf{x})}{\pi_0 f_0(\mathbf{x}) + \pi_1 f_1(\mathbf{x})} = \frac{1}{1 + e^z} \quad (3)$$

por lo que se arriba a que el logaritmo del cociente de probabilidades es lineal,

$$\ln \frac{pr(Y = 1 / \mathbf{x})}{pr(Y = 0 / \mathbf{x})} = z \quad \text{siendo } z = \beta_0 + \beta' \mathbf{x} \quad (4)$$

Para interpretar los coeficientes en una regresión logística es necesario definir la ecuación obtenida de una manera más conveniente, expresando el modelo en términos de las chances de que un evento ocurra, y ello se obtiene exponenciando la expresión (4).

$$\frac{pr(Y = 1 / \mathbf{x})}{pr(Y = 0 / \mathbf{x})} = e^z \quad (5)$$

De esta manera, las chances que un evento ocurra (odds) quedan definidas como el cociente entre la probabilidad de que el evento ocurra ($Y = 1$) y la probabilidad de que no ocurra ($Y = 0$). Los coeficientes estimados se interpretan como el cambio en el logaritmo del odds por un cambio unitario en la variable independiente asociada, permaneciendo las demás variables constantes. El cociente de odds para una variable independiente dicotómica, indica en cuánto aumenta (o disminuye) la chance que se presente el evento entre los individuos que pertenecen a una categoría, en relación con la categoría de referencia. Este cociente se puede extender a variables independientes politómicas tomando una de las categorías como referencia.

En el modelo logístico con g grupos, se elige uno de ellos como base, razón por la cual la expresión (1), considerando como base el g -ésimo grupo, se generaliza de la siguiente manera:

$$\ln \left\{ \frac{f_j(\mathbf{x})}{f_g(\mathbf{x})} \right\} = \beta_{0j} + \beta_j' \mathbf{x} \quad j = 1, 2, \dots, g-1 \quad (6)$$

La probabilidad a posteriori para cada uno de los grupos se obtiene haciendo

$$pr(Y = j / \mathbf{x}) = \frac{e^{z_j}}{\sum_{j=0}^{g-1} e^{z_j}} \quad j = 1, 2, \dots, g-1 \quad \text{donde } z_0 = 0 \quad (7)$$

Si la variable respuesta asume g valores y es de naturaleza ordinal, se puede aplicar el *Modelo de Regresión Ordinal*, en el que se establece una de las categorías de Y como grupo base (generalmente la última) y se obtienen $g-1$ funciones que difieren sólo en la constante. La naturaleza ordinal de la respuesta permite obtener las probabilidades acumuladas representadas por γ_j de conformidad a la siguiente tabla.

Categoría	1	2	...	g
Prob. Categoría	$\pi_1 = \text{pr}(Y=1)$	$\pi_2 = \text{pr}(Y=2)$...	$\pi_g = \text{pr}(Y=g)$
Prob. Acumulada	$\gamma_1 = \text{pr}(Y=1)$	$\gamma_2 = \text{pr}(Y \leq 2)$...	$\gamma_g = 1$

En el modelo propuesto por Mc Cullagh y Nelder (1989), la expresión del logit (4) para las probabilidades acumuladas es la siguiente:

$$\text{Log} \frac{\gamma_j}{1-\gamma_j} = \beta_{0j} - \beta' \mathbf{x} \quad (8)$$

Esta expresión supone que existe una variable² no observable L , con distribución logística y media $\beta' \mathbf{x}$, relacionada con la respuesta de acuerdo a la siguiente regla:

$$Y = j \text{ si } \beta_{0(j-1)} < L \leq \beta_{0j} \quad j = 1, 2, \dots, g-1 \quad (9)$$

Así, integrando las relaciones introducidas en el modelo ordinal, la probabilidad acumulada a posteriori $\gamma_j = \text{pr}(Y \leq j) = \text{pr}(L \leq \beta_{0j})$ se puede expresar como:

$$\gamma_j = \frac{e^{z_j}}{1+e^{z_j}}, \quad \text{siendo } z_j = \beta_{0j} - \beta' \mathbf{x} \quad (10)$$

Tal como se advierte en la expresión de z_j , un coeficiente positivo disminuye la probabilidad de la primera categoría, aumentando la de la categoría de referencia.

Para probar si es correcto el supuesto de líneas paralelas se utiliza el estadístico G , que supone pendientes iguales a título de hipótesis nula ($H_0: \beta_1 = \beta_2$). En esta prueba se compara la verosimilitud obtenida bajo el modelo restringido por H_0 (coeficientes iguales) con la que resulta en un modelo no restringido (modelo Multinomial): $-2 \ln L_0 - 2 \ln L_1 : \chi^2$. Bajo hipótesis nula, el estadístico Chí-Cuadrado se distribuye con $(g-2)p$ grados de libertad.

²La variable latente es denominada Tolerancia en aplicaciones biológicas y Utilidad en economía.

III.2. Árboles de Decisión

En la actualidad se advierte un proceso sostenido de aceptación y utilización de los denominados Árboles de Decisión, sobre todo en aquellos problemas donde se trata de particionar conjuntos de datos e identificar sus estructuras subyacentes. También han quedado comprobadas sus ventajas en problemas de clasificación supervisada, ya se trate de pequeños o grandes conjuntos de datos.

En la operatoria de estos modelos se entiende que la clasificación es una partición del espacio de mediciones X (que incluye todo el conjunto posible de estados del vector de predictores) en J subconjuntos disjuntos.

El algoritmo aplicado al presente conjunto de datos, CHAID (Detección de interacción automática Chi-cuadrado) es una técnica estadística muy eficiente para segmentar o generar árboles. Presentada y desarrollada por Kass (1980), utiliza como criterio la significación de un contraste estadístico, evaluando todos los valores de una variable predictora potencial. Reúne además los valores considerados estadísticamente homogéneos respecto de la variable criterio y conserva inalterados todos los valores que resulten distintos (no homogéneos). Luego selecciona la mejor variable predictora para formar la primera rama del árbol de decisión, de forma que cada nodo esté compuesto por un grupo de valores homogéneos de la variable seleccionada y así se continúa hasta desarrollar por completo el árbol. La prueba estadística utilizada cuando la variable criterio es categórica es la Chi-cuadrado. Este algoritmo incorpora un procedimiento secuencial de reunión de los individuos que conforman la base de datos cuando un par de categorías del predictor (tabla $2 \times k$) no presenta asociación con la variable criterio y de división cuando están relacionadas.

III.3. Estimación de la tasa de error

Las reglas de decisión o las funciones clasificadoras derivadas a partir de los datos, permiten obtener una estimación de la probabilidad condicional de cada clase en particular, fijado el vector de mediciones que corresponde a un elemento.

Si sólo hay dos clases, el espacio se particiona en dos regiones mutuamente excluyentes R_1 y R_2 . Si $x \in R_1$ la observación se clasifica como G_1 y en caso contrario se clasifica como G_2 . En cualquier caso, siempre se puede cometer el error de asignación a una categoría incorrecta.

Entre las estimaciones de la tasa de error se encuentra la *Tasa de error aparente*, que consiste en establecer la proporción de objetos de la muestra que fueron mal clasificados. Esta tasa subestima considerablemente la verdadera *tasa de error condicional*, ya que reclasifica los elementos de la muestra con la regla determinada a partir de ellos mismos.

Una forma de corregir el sesgo de la tasa aparente consiste en aplicar el denominado Método Holdout, una técnica que permite probar el modelo con otro conjunto de datos. Para ello se subdividen los datos disponibles en dos partes independientes: una de ellas (*muestra de entrenamiento*) se utiliza para desarrollar la regla de clasificación que es aplicada al segundo grupo, generalmente designado como *muestra test*. Al obtener la proporción de objetos mal clasificados, permite verificar la precisión del modelo desarrollado con la muestra de análisis.

Otra alternativa es calcular la llamada tasa *cross-validada*, que se obtiene siguiendo los siguientes pasos: 1) se omite cada vez una observación del conjunto de datos; 2) recálculo de la regla de predicción sobre la base de las restantes n-1 observaciones; 3) utilización de la regla para clasificar la observación omitida y 4) se calcula la proporción de errores que se cometieron con este procedimiento.

IV. RESULTADOS OBTENIDOS

En esta sección se analizan detallada y críticamente los resultados obtenidos con los modelos estimados aplicando Regresión Logística y Árboles de Decisión. Las otras técnicas utilizadas, Vecino más Cercano y Redes Neuronales, sólo permiten estimar las tasas de error, las que se presentan en la sección V conjuntamente con la comparación de la performance alcanzada por todos los métodos ³.

La variable respuesta es la condición de actividad, que distingue Ocupados, Subocupados y Desocupados. En la Tabla 1 se presentan las variables predictoras seleccionadas en el análisis.

³ Para el procesamiento se utilizaron los siguientes paquetes estadísticos: Regresión Logística: paquete estadístico SPSS 11.5, Árboles de decisión: paquete estadístico Answer Tree de SPSS, Redes Neuronales: subrutinas del paquete matemático MATLAB, Vecino más cercano: paquete estadístico SAS.

Tabla 1
Variables Predictoras Seleccionadas

<i>A) Variables Numéricas</i>	
Edad	años cumplidos
Personas en el hogar	Cantidad de habitantes en el hogar
<i>B) Variables Categóricas</i>	
Sexo	1- Mujer 2- Varón
Estado civil	1- Sin pareja 2- Con pareja
Parentesco	1- No Jefe 2- Jefe
Nivel de instrucción	1- Sin educación - Primario incompleto 2- Primario completo - Secundario incompleto 3- Secundario completo - Terciario incompleto 4- Terciario completo
Localidad	1- Córdoba 2- Gran Buenos Aires 3- Rosario
Categoría ocupacional	1- Patrón o empleador 2- Trabajador por su cuenta 3- Obrero o empleado
Condición NBI	1- Con NBI 2- Sin NBI
Rama de actividad	1- Industria 2- Construcción 3- Comercio 4- Servicios comunales - Administración. Pública 5- Otras ramas

IV.1. Resultados de la Regresión Logística

Una vez analizadas la pertinencia de las variables a incluir en la regresión logística y formalizado un minucioso estudio previo de sus contribuciones individuales, se prosiguió con la construcción de las escalas

para las variables intervalares (*Edad* y *Personas en el hogar*) y el agrupamiento de categorías adyacentes para las nominales. A este respecto, es oportuno señalar que todas las variables, excepto *Localidad*, están asociadas con la condición de actividad, razón por la cual se optó por excluir esta última del modelo multivariado.

Para la variable *Edad* se consideró que el agrupamiento apropiado es dicotomizar sus valores en dos categorías: a) hasta 26 años y b) más de 26 años, en tanto que para la otra variable numérica (*Personas en el hogar*) el agrupamiento más conveniente resultó: a) hogares con 3 personas o menos, b) hogares con 4 o 5 personas y c) hogares con 6 o más personas.

En relación con las variables categóricas, la única que se modifica para la obtención del modelo es *Categoría Ocupacional*, agrupando las dos primeras categorías. De ese modo, se utilizó una variable con dos modalidades: a) trabajador independiente y b) trabajador en relación de dependencia.

En una primera etapa se adaptó el Modelo Logístico Ordinal, que estima dos funciones que sólo difieren en la constante. No es oportuno detenerse en el análisis de los resultados ya que se rechaza la hipótesis de líneas paralelas, que supone igualdad de coeficientes de los predictores de las dos funciones binarias, esto es: 1) Desocupado versus Ocupados y 2) Subocupados versus Ocupados, resultando apropiado la aplicación del Modelo Logístico Multinomial.

Asimismo, se consideró adicionalmente la incorporación de algunas interacciones juzgadas pertinentes en esta aplicación, tales como *Edad* y *Jefatura de hogar* (Int. 1), *Edad* y *Sexo* (Int. 2), *Sexo* y *Estado Civil* (Int. 3), *Estado Civil* y *Jefatura de hogar* (Int. 4), *Edad* y *Estado Civil* (Int. 5) y *Jefatura de hogar* y *Sexo* (Int. 6). Conforme a los resultados consignados en la Tabla 2, donde se muestran los resultados de los test de la razón de verosimilitud (G) sólo las dos primeras resultaron significativas ($p < 0.05$) mas no así las restantes. Cabe apuntar que estas pruebas estadísticas comparan el modelo con cada interacción versus el modelo que contiene sólo los efectos principales (M1). El modelo que incluye las dos interacciones es comparado primero con M1 siendo su diferencia significativa, y luego con M2, que incluye sólo la primera interacción, resultando en este caso no significativa. En conclusión, como modelo final se adopta el que contiene los efectos principales y la interacción *Edad* y *Jefatura de hogar*.

Tabla 2
Test de la Razón de Verosimilitud (G) para Interacciones Significativas

Cód.	Modelo	-2 log de la verosimilitud	Modelo de referencia	G	gl	Sig.
	Términos					
M1	Efectos Principales	3642.25	M1			
M2	M1 más Int. 1	3632.14	M1	10.1	2	0.006
M3	M1 más Int. 2	3635.64	M1	6.61	2	0.037
M4	M1 más Int.1 e Int.2	3628.52	M1	13.73	4	0.008
			M2	3.62	2	0.163

En la Tabla 3 se muestran los cocientes de odds y los intervalos de confianza de los coeficientes que resultaron significativos.

Tabla 3
Cocientes de Odds e Intervalos de Confianza para los parámetros significativos

Condición de actividad	Exp(β)	Intervalo de confianza al 95% para Exp(β)	
		Límite inferior	Límite superior
Desocupados	[SEXO=Mujer]	1.670	1.359 2.054
	[ESTADO CIVIL=Sin pareja]	1.917	1.568 2.345
	[JEF. DE HOGAR=No jefe]	1.575	1.252 1.981
	[EDUCACION=Prim.Incompleto]	2.634	1.725 4.022
	[EDUCACION=Sec. Incompleto]	2.236	1.657 3.018
	[RAMA=Construcción]	2.730	2.007 3.714
	[RAMA= Adm. Pública]	.654	.501 .853
	[NBI=Si]	2.036	1.673 2.477
	[PERS.EN EL HOGAR=4 o 5]	1.350	1.061 1.719
	[CAT. OCUPACIONAL=Indep.]	1.316	1.066 1.623
	[J. HOGAR=No jefe]*[EDAD=H 26]	1.968	1.112 3.482
Subocupados	[SEXO= Mujer]	2.167	1.768 2.656
	[ESTADO CIVIL= Sin pareja]	1.476	1.215 1.792
	[JEF. DE HOGAR= No jefe]	1.463	1.173 1.825
	[EDUCACION=Prim.Incompleto]	6.160	4.235 8.961
	[EDUCACION=Sec. Incompleto]	3.154	2.395 4.152
	[RAMA=Industria]	.618	.457 .834
	[RAMA=Comercio]	.625	.489 .798
	[RAMA= Adm. Pública]	1.960	1.564 2.455
	[NBI=Si]	1.359	1.119 1.652
	[EDAD=Hasta 26]	1.844	1.224 2.779
	[PERS.EN EL HOGAR=Hasta 3]	.763	.595 .979
	[CAT. OCUPACIONAL= Indep.]	2.291	1.876 2.799

IV.2. Factores condicionantes del desempleo

Basado en las características personales de los individuos y bajo el supuesto de permanencia de los demás factores interactuantes, el análisis de los coeficientes estimados en el modelo de regresión muestra que las chances de desocupación son un 67% más alta en mujeres que en los varones. En relación con el *nivel de educación* no se observan diferencias significativas entre quienes sólo han completado el secundario y los que alcanzaron el título terciario/universitario. Sin embargo, al descender en el nivel de educación formal, el cociente de chances se incrementa notablemente para aquellos que sólo han alcanzado un nivel de estudios inferior a secundario completo. Así, en la comparación de las categorías "primario incompleto" y "terciario completo", ese cociente es igual a 2.6, relación que disminuye levemente al contrastar "hasta secundario incompleto" (cociente de chances igual a 2.2).

Por sí misma, la edad no tiene un peso significativo en la desocupación, a menos que se la tome en cuenta asociada a la *condición de parentesco*. El no ser jefe de hogar es un factor de riesgo de desocupación en los dos grupos de edad, mientras que la edad menor a 26 años aparece como una dificultad para encontrar empleo entre quienes no son jefes de hogar.

Desde otra perspectiva y por lo que se refiere al *estado civil*, las personas sin pareja tienen un 90% más de chances de desempleo que aquellos que tienen pareja. Asimismo, quienes viven en hogares de tamaño mediano (4 a 5 personas), tienen mayor probabilidad de estar desocupados que quienes habitan en hogares con más de 6 componentes. Finalmente, resulta perceptible el efecto negativo de vivir en un *hogar con NBI*, toda vez que ello aumenta un 100% las chances de desocupación.

Con respecto a los indicadores de la historia laboral, y específicamente en lo atinente a la *rama de ocupación* (para el último empleo en el caso de desocupados), la mayor probabilidad de ser desocupado la tienen quienes trabajaban en la rama de la construcción. Por el contrario, el trabajar en la administración pública reduce la frecuencia de desocupación. El otro indicador de la historia laboral es la *categoría ocupacional*, y tal como era de esperar, el ser trabajador independiente aumenta las chances de desempleo en un 30%, con respecto a los que trabajan en relación de dependencia.

Jerarquizando el efecto de los determinantes del desempleo, resulta claro que los de mayor impacto lo constituyen el haber trabajado en la rama

de la construcción, en segundo lugar el nivel educativo, seguido luego por la condición de vivir en hogares con NBI, no tener pareja, no ser jefe de hogar y tener menos de 26 años.

El análisis anterior puede efectuarse también calculando las probabilidades de desocupación según distintas características socio-económicas de las personas. Esas probabilidades se muestran en la tabla siguiente, en la cual el individuo definido como referencia tiene estudios terciarios completos, vive con pareja en un hogar sin NBI, es jefe del hogar y trabajar en una ocupación ubicada en la categoría otras ramas de actividad. A partir de ese patrón, se cuantificó el incremento en las probabilidades del desempleo a medida que se incorporan sucesivamente los factores de riesgo de desempleo.

Tabla 4
Probabilidades de desocupación por sexo

	Varón	Mujer
Individuo de referencia	0.06	0.10
Primario Incompleto	0.15	0.23
Hogar con NBI	0.27	0.38
Sin pareja	0.41	0.38
No Jefe de hogar	0.53	0.65
Rama Construcción	0.75	

IV.3. Factores condicionantes del subempleo

El modelo desarrollado para esta partición de la muestra señala que el bajo nivel de educación es el factor que más impacto tiene en el subempleo, toda vez que la frecuencia de subempleo es 6 veces más alta entre personas con estudios primarios incompletos que entre aquellos que han completado el nivel terciario. A su vez, para aquellos que realizaron estudios secundarios (aunque sin completarlos), las chances descienden a la mitad.

Asimismo, el ser mujer o tener menos de 26 años también constituyen factores de riesgo. A diferencia del modelo anterior, la *edad* considerada como efecto principal juega un papel importante, ya que entre los menores de 26 años se duplica el riesgo de subocupación.

Cabe consignar que en este modelo particular disminuyen la incidencias de las características del hogar, con efectos similares para *estado civil*, *jefatura de hogar* y *hogares con NBI*. También es diferente la incidencia de

la *rama de actividad*, ya que se manifiesta una fuerte subocupación en las personas que trabajan en la administración pública, en tanto que pertenecer al sector Industria y Comercio reduce la chance de subempleo. La posición trabajador independiente señala un mayor efecto en el subempleo que en el desempleo.

Ordenando el efecto de los determinantes del subempleo de acuerdo a su importancia, en primer lugar se destaca el nivel educativo, seguido por el hecho de ser trabajador independiente, ser mujer y trabajar en la administración pública.

IV.4. Resultados del método de árboles de decisión

Tomando las escalas para las variables intervalares y nominales descritas en el punto anterior, el árbol resultante con la variable criterio Multinomial (tres categorías) no fue eficiente, ya que arroja elevados errores de clasificación. En una segunda etapa del análisis, se desarrollaron dos árboles, a saber: Desocupados/Ocupados y Subocupados/Ocupados, cuyos análisis se exponen a en los puntos siguientes.

IV. 4.1. Árbol resultante del análisis Ocupados-Desocupados (Figura 1 - Anexo)

Conforme a modalidades que son propias de este método, la primera división del árbol se realiza a partir de la variable *condición de parentesco*, de tal manera que la categoría no jefe de hogar aparece con mayores probabilidades de desocupación (37%).

Asimismo, los jefes de hogar se segmentan según la *rama de actividad* en la que trabajan, y se observa que quienes lo hacen en industria, comercio o servicios comunales y sociales exhiben mayores posibilidades de aparecer ocupados (81% y 89% respectivamente). Inversamente, los jefes de hogar que trabajan en el sector construcción muestran un elevadísimo porcentaje de desocupación, que llega al valor del 50%.

En una instancia ulterior de segregación, los que trabajan en la construcción y en el área de servicios comunales se clasifican según el *nivel de educación*. Para los primeros, los que ostentan niveles de educación en los extremos, es decir, altos y bajos, tienen mayor probabilidad de estar ocupados, ya que unos se desempeñarán en los niveles de dirección o profesional y los otros en las tareas menos calificadas, mientras que los que poseen una formación media tienen más probabilidad de estar desocupados en esta rama

de actividad (59%). Para los que trabajan en el área de servicios comunales y sociales, el porcentaje de ocupación es alto para cualquier nivel de educación, destacando que quienes alcanzaron un nivel alto de capacitación están ocupados en un 94%. En la industria y el comercio los miembros de la PEA se clasifican según el sexo, apareciendo con mayores probabilidades de desocupación las mujeres (35%).

Para el grupo de los No Jefes, el árbol tiene dos brazos, a saber: a) los que tienen NBI y b) los que no tienen dicha condición. Los primeros se clasifican según el número de personas en el hogar, siendo los que pertenecen a hogares más numerosos los que tienen más probabilidad de estar desocupados. Por su parte, los que no tienen NBI se dividen según el *estado civil*, pudiéndose advertir que quienes viven en pareja presentan un mayor porcentaje de ocupación (89%), lo cual permite confirmar la hipótesis de que estas personas realizan esfuerzos mayores para lograr una ocupación.

En resumen, se detectaron 11 clases (Tabla 5) ordenadas por riesgo decreciente de desocupación y en la Tabla 5 se muestran los nodos ordenados según la proporción de desocupados.

Tabla 5
Clases en el Análisis Ocupados - Desocupados

Clase 1	No Jefes con NBI con 4 o 5 personas en el hogar
Clase 2	Jefes que trabajan en la construcción y que tienen nivel de instrucción medio
Clase 3	No Jefes con NBI con 6 o más personas en el hogar
Clase 4	Jefes que trabajan en la construcción y que tienen nivel de instrucción bajo o alto
Clase 5	No Jefes sin NBI y sin pareja
Clase 6	Jefes Mujer que trabajan en la industria, comercio u otra rama
Clase 7	No Jefes con NBI con 1 a 3 personas en el hogar
Clase 8	No Jefes sin NBI y con pareja
Clase 9	Jefes que trabajan en servicios comunales y sociales y que tienen nivel de instrucción bajo o medio
Clase 10	Jefes Varón que trabajan en la industria, comercio u otra rama
Clase 11	Jefes que trabajan en servicios comunales y sociales y que tienen nivel de instrucción alto

Tabla 6
Cuadro de Ganancia en el Análisis Ocupado - Desocupado

Clase	Tamaño del nodo	% que representa cada nodo	Cantidad de desocupados en cada nodo	Distr. de desocupados en %	% que representa los desocupados en el nodo	Índice*
Clase 1	169	5,82	109	13,13	64,50	225,58
Clase 2	108	3,72	64	7,71	59,26	207,26
Clase 3	178	6,13	95	11,45	53,37	186,67
Clase 4	77	2,65	29	3,49	37,66	131,73
Clase 5	549	18,91	195	23,49	35,52	124,23
Clase 6	136	4,68	48	5,78	35,29	123,44
Clase 7	57	1,96	20	2,41	35,09	122,72
Clase 8	382	13,16	76	9,16	19,90	69,59
Clase 9	133	4,58	24	2,89	18,05	63,11
Clase 10	934	32,17	159	19,16	17,02	59,54
Clase 11	180	6,20	11	1,33	6,11	21,37
	3013	100,00	940	100,00		

**relación de % de desocupados en cada nodo respecto al % en el total de la muestra*

Analizando el cuadro anterior, se advierte que cuando el índice presentado en la última columna supera el 200%, tal como ocurre en las clases 1 y 2, la proporción de desocupados duplica el porcentaje de esa categoría en la muestra total. Es decir, que las clases "No jefes con NBI, con 4 o 5 personas en el hogar" (Índice 225,58), "Jefes que trabajan en la construcción y que tienen nivel de instrucción medio" (Índice 207,26) representan los grupos de mayor riesgo. En el otro extremo, los "Jefes que trabajan en servicios comunales y sociales y que tienen nivel de instrucción alto", que componen la clase 11, corresponden al grupo de menor riesgo, con un porcentaje de desempleo del 6,11%, el que representa sólo un quinto del valor de la muestra total.

IV.4.2. *Árbol resultante del análisis Ocupados-Subocupados (Figura 2 - Anexo)*

En este segundo esquema clasificatorio, la primera división del árbol se realiza a partir de la variable sexo, algo que en principio estaría sugiriendo que las mujeres tienen más probabilidad de estar subempleadas que los hombres. A su vez, los varones se clasifican según su *categoría ocupacional* y allí se puede observar que para el grupo patrones o empleadores se ha alcanzado lo que se denomina "pureza", lo que convierte a este segmento demográfico en nodo terminal. En este grupo, prácticamente la totalidad de sus miembros son ocupados, mientras que los nodos que identifican a los trabajadores cuentapropistas y obrero o empleado pueden todavía aceptar otras divisiones.

Los cuentapropistas se segmentan según la *edad*, siendo los más jóvenes los que tienen mayores probabilidades de estar subocupados. Por otra parte, los que tienen una ocupación en relación de dependencia se clasifican por la *rama de actividad* de la empresa en la que trabajan. Los agrupados en esta última segmentación presentan las más altas probabilidades de empleo y por consiguiente mayores horizontes de estabilidad en la ocupación.

Analizando la rama del árbol que corresponde a las mujeres, la condición de ocupación o subocupación depende del *nivel de instrucción*. En tal sentido, las que presentan menor nivel educacional tienen mayores probabilidades de estar subocupadas, constituyéndose de este modo en un nodo terminal.

Por otra parte, las que poseen mejor formación educativa pueden clasificarse a su vez según la *rama de actividad*, pero en casi todas las ramas las mujeres con nivel de instrucción alto tienen más posibilidades de estar ocupadas, es decir, no necesitan buscar otro trabajo, siendo las únicas excepciones aquéllas que han alcanzado el nivel medio y trabajan en servicios comunales, grupo para el que el nivel de subocupación es del 67%. La Tabla 7 resume las clases que se detectaron en el análisis y en la Tabla 8 se ordenan los nodos según la proporción de subocupados, es decir, permite saber qué subgrupo de personas tiene mayores probabilidades de ser clasificado como tal.

Tabla 7
Clases en el Análisis Ocupado - Subocupado

Clase 1	Mujeres sin instrucción o primario incompleto
Clase 2	Mujeres con primario completo o secundario incompleto que trabajan en servicios comunales o sociales u otra rama
Clase 3	Varones que trabajan por su cuenta menor de 26 años
Clase 4	Mujeres con primario completo o secundario incompleto que trabajan en la industria, construcción o comercio
Clase 5	Mujeres con alto nivel de instrucción que trabajan en servicios comunales o sociales
Clase 6	Varones que trabajan por su cuenta de 26 años o más
Clase 7	Varones obrero o empleado que trabajan en construcción o servicios comunales o sociales
Clase 8	Mujeres con alto nivel de instrucción que trabajan en la industria, construcción, comercio u otra rama
Clase 9	Varones obrero o empleado que trabajan en comercio
Clase 10	Varones obrero o empleado que trabajan en industria u otra rama
Clase 11	Varones con categoría ocupacional patrón u empleador

Tabla 8
Cuadro de Ganancia en el Análisis Ocupado - Subocupado

Clase	Tamaño del nodo	% que representa cada nodo	Cantidad de subocupados en cada nodo	Distr. de subocupados en %	% que representa los subocupados en el nodo	Índice*
Clase 1	82	2,72	61	6,49	74,39	238,44
Clase 2	254	8,43	169	17,98	66,54	213,27
Clase 3	72	2,39	47	5,00	65,28	209,24
Clase 4	128	4,25	53	5,64	41,41	132,72
Clase 5	288	9,56	108	11,49	37,50	120,20
Clase 6	444	14,74	159	16,91	35,81	114,79
Clase 7	419	13,91	136	14,47	32,46	104,04
Clase 8	357	11,85	73	7,77	20,45	65,54
Clase 9	268	8,89	50	5,32	18,66	59,80
Clase 10	620	20,58	79	8,40	12,74	40,84
Clase 11	81	2,69	5	0,53	6,17	19,79
	3013	100,00	940	100,00		

*relación de % de subocupados en cada nodo respecto al % en el total de la muestra

Realizando un razonamiento similar al efectuado en el análisis ocupados-desocupados, se advierte que las clases 1 a 3 duplican el porcentaje de subocupados en relación con la muestra total, es decir que "Mujeres sin instrucción o primario incompleto", "Mujeres con primario completo o secundario incompleto que trabajan en servicios comunales o sociales u otra rama" y "Varones menores de 26 años que trabajan por su cuenta" constituyen los grupos de mayor riesgo. Por el contrario "Varones con categoría ocupacional patrón u empleador" tienen muy poca probabilidad (6,17%) de ser subocupados.

V. COMPARACIÓN DE LAS TASAS DE ERROR

Los otros dos métodos aplicados -Redes Neuronales y Vecino más Cercano⁴ - tienen la importante desventaja de no permitir identificar los factores que tienen incidencia sobre el empleo, pero su aplicación en este trabajo tuvo por objetivo verificar si es posible mejorar la separación de los grupos definidos por la variable criterio. En la literatura estadística se sugiere el modelo de Redes Neuronales como un método capaz de particionar el espacio de medición en regiones de una manera más eficiente que el Modelo Logístico, por lo que es dable esperar una reducción de la tasa de error⁵ y entre los métodos no paramétricos, el del Vecino más Cercano es la opción correcta cuando las variables predictoras son categóricas.

En los modelos Logístico y Redes Neuronales, una vez estimados los coeficientes, se obtienen las probabilidades *a posteriori*, a partir de cuyos valores se asignan los casos a uno de los dos grupos bajo análisis. En la mayoría de las aplicaciones se fija el umbral en 0,5, de manera que para que un individuo sea asignado al grupo Ocupado, su probabilidad estimada debe ser mayor que ese valor. En este trabajo se adoptó el punto de corte 0,7, ya que ello reduce la cantidad de Desocupados y Subocupados mal clasifica-

⁴ El Método del Vecino más Cercano no permite realizar la estimación de parámetros, mientras que el Modelo de Redes Neuronales los estima pero no son interpretables, por lo que con ambos métodos no es posible describir los factores que fueron determinantes en la clasificación de los individuos en cada grupo.

⁵ En la aplicación de Redes Neuronales se probaron distintas tipologías con diferentes funciones de activación, logrando en todos los modelos mejores resultados con una capa oculta con función de activación tangente y la capa de salida con función de activación logística.

dos, aún cuando este criterio arroje una mayor tasa global⁶. Este inconveniente no se presentó en el modelo del Vecino más Cercano, en el que se mantuvo el punto de corte de 0,5.

La tasa de error en el Modelo Logístico, cuando se fija el punto de corte en 0,5, es muy similar a la obtenida en los Árboles de Decisión, toda vez que ambos métodos son muy sensibles a los tamaños de los grupos y por consiguiente tienden a favorecer la asignación al grupo de mayor tamaño, es decir Ocupados para este conjunto de datos. Sin embargo dicho inconveniente que pudo ser subsanado en el Modelo Logístico al requerir una alta probabilidad para asignar los casos a ese grupo, no se puede aplicar en los Árboles de Decisión.

Dejando de lado los resultados alcanzados con Árboles de Decisión, en la Tabla 9 se resumen los porcentajes de error obtenidos en las clasificaciones binarias, donde se destaca claramente que las proporciones de casos mal clasificados, en general se mantienen para cada método en ambas categorías.

Tabla 9
Resumen de las tasas de error aparente resultantes en la discriminación binaria

Modelo	Desoc.	Suboc.	Ocup.	Global
<i>Desocupados-Ocupados</i>				
Logístico	0.37		0.30	0.32
Redes Neuronales	0.33		0.25	0.27
Vecino más cercano Uniforme	0.21		0.19	0.19
<i>Subocupados-Ocupados</i>				
Logístico		0.30	0.35	0.33
Redes Neuronales		0.32	0.28	0.29
Vecino más cercano Uniforme		0.21	0.22	0.22

⁶Tasas de error aparente resultantes en la discriminación binaria con punto de corte de 0.5

Modelo	Desoc.	Suboc.	Ocup.	Global
<i>Desocupados-Ocupados</i>				
Logístico	0.74		0.07	0.26
Redes Neuronales	0.64		0.06	0.23
<i>Subocupados-Ocupados</i>				
Logístico		0.66	0.09	0.25
Redes Neuronales		0.58	0.08	0.23

El Vecino más Cercano es el que logró una mejor performance, con tasas de alrededor del 20%, ubicándose en el otro extremo el Modelo Logístico con valores que superan el 30%. Si se compara este modelo con el de Redes Neuronales, se advierte una leve mejora de los resultados que no justifica la utilización de esta técnica, dada su compleja interpretación.

La subestimación de la tasa aparente fue medida para el modelo del Vecino más Cercano, en el que resultaron las siguientes tasas de error crossvalidadas: de 0,35 para los grupos Ocupados vs. Desocupados y de 0,38 para Ocupados vs. Subocupados.

VI. CONCLUSIONES

El análisis de los resultados de la Regresión Logística y de Árboles de Decisión, permite concluir que los factores explicativos de la probabilidad de que una persona sea desocupada o subocupada concuerdan en general con las hipótesis planteadas.

Respecto a la Regresión Logística se obtuvieron dos funciones: 1) Desocupados versus Ocupados y 2) Subocupados versus Ocupados. No obstante la disimilitud de los coeficientes de los factores, se observa que las características socio-demográficas y laborales que inciden en los dos grupos de riesgo laboral son las mismas, algunas con mayor incidencia en desocupación y otras en subocupación.

Es importante destacar que estos modelos permiten identificar los factores prevalentes que inciden en la problemática, siempre que se los analice desde una perspectiva multidimensional, esto es, considerando simultáneamente todas las variables predictoras.

En cambio, en los nodos terminales del Árbol de Decisión, se pueden identificar grupos de casos, ordenados según el porcentaje que presentan para la variable criterio. Precisamente, esta es la mayor contribución de esta metodología, ya que la lectura de las diferentes ramas -correspondientes a esos nodos- brinda la caracterización de los perfiles de cada una de las clases.

Los otros dos métodos aplicados -Redes Neuronales y Vecino más Cercano- tienen la desventaja de no permitir identificar los factores que tienen incidencia sobre la probabilidad de que una persona tenga problemas de empleo. Su aplicación en este trabajo no va más allá del objetivo de verificar si es posible mejorar la separación de los grupos definidos por la variable criterio.

VII. REFERENCIAS

- Aitchison, J. y Aitken, C. (1976). "Multivariate binary discrimination by the kernel method". *Biometrika*, 63, 413-420.
- Anderson, T.W. (1984, 2da.edic.). *An Introduction to Multivariate Statistical Analysis*. 1ra edic. (1958). New York, Wiley.
- Biggs, D. B. de Ville, and Suen, E. (1991). "A method of choosing multiway partitions for classification and decision trees". *Journal of Applied Statistics*, 18, 49-62.
- Breiman, L.Friedman, J. Olshen, R. y Stone, C. (1998) *Classification and regression Trees*. New York, Chapman & Hall.
- Diaz, Margarita (2001). Performance del Análisis Discriminante Regularizado en la predicción de Crisis Financieras. *Tesis Maestría en Estadística Aplicada*. UN de Córdoba.
- Hall, P. y Matthew, W. (1988). "On Nonparametric Discrimination Using Density Differences". *Biometrika*, vol. 75, N° 3, 541-547.
- Hand, D.J. (1981). *Discrimination and Classification*. New York, Wiley.
- _____ (1982). *Kernel discriminant analysis*. New York, Wiley.
- _____ (1999). *Construction and Assessment of Classification Rules*. Chichester, Wiley.
- Hills, M. (1967). "Discrimination and allocation with discrete data". *Applied Statistics*, 16, 237-250.
- Hosmer, D. and Lemeshow, S. (1989). *Applied Logistic Regression*. New York, Wiley.
- <http://www.indec.mecon.gov.ar/eph/eph03.asp>
- <http://www.indec.gov.ar/nuevaweb/cuadros/4/metempleo1.pdf>
- Izenman, Alan (1991). "Recent Developments in Nonparametric Density Estimation". *JASA*, Vol. 86 N° 413, 205-224
- Johnson, R.A. y Wichern, D.W. (1992, 3 edic.). *Applied Multivariate Statistical Analysis*. 1ra edic.(1982). New York, Prentice-Hall
- Kass, G (1980). "An exploratory technique for investigating large quantities of categorical data". *Applied Statistics*, 29:2, 119-127.
- McCullagh P. and Nelder JA (1989). *Generalized Linear Models*. London, Chapman and Hall.
- McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York, Wiley.
- SPSS Inc. *Manual del Usuario "Answer Tree 2.0."* (1998).
- Wang, C y Van Ryzin, J (1981). "A class of smooth estimators for discrete distributions". *Biometrika* 68, 301-309.
- Wilkinson, Leland (1992). "Tree Structured Data Analysis: AID, CHAID and CART", *SPSS Inc.*, Chicago, Department of Statistics, Northwestern University.