Reportes Científicos de la FACEN

Reportes científicos de la FACEN, Julio - Diciembre 2022, 13(2):122–130

http://doi.org/10.18004/rcfacen.2022.13.2.122

Artículo Original

Estudio comparativo por métodos de clasificación para el análisis del desempleo en los departamentos de la región oriental del Paraguay

Comparative study by classification methods for the analysis of unemployment in the departments of the eastern region of Paraguay

Lorena Leticia González[1,*], Luis Antonio Gómez Martínez[1,2] & María Cristina Martín[3,4]

[1]Universidad Nacional de Asunción, Facultad de Ciencias Exactas y Naturales, Departamento de Estadística, San Lorenzo, Paraguay.
[2]Universidad Nacional de Asunción, Facultad de Filosofía, Asunción, Paraguay.
[3]Universidad Nacional de La Pampa, Facultad de Ciencias Exactas y Naturales, Santa Rosa, La Pampa, Argentina.
[4]Universidad Nacional del Sur, Departamento de Matemática, Bahía Blanca, Buenos Aires, Argentina.
*Autor correspondiente: lorenleti@gmail.com.

Resumen: El problema de la clasificación de individuos u objetos en grupos o poblaciones conocidas es de gran interés en estadística, por esta razón se han desarrollado varias técnicas para cumplir este propósito. Este trabajo propende a identificar los factores de riesgo que inciden en la precariedad laboral de la población paraguaya, adoptando como base de información la Encuesta Permanente de Hogares 2011. El efecto de las variables predictoras (edad, sexo, estado civil, nivel de educación, parentesco con el jefe de hogar, departamento, área, rama y categoría del último empleo) sobre la Situación Laboral del encuestado se estima a través de los Análisis de Regresión Logística y Árboles de Clasificación. El análisis de los resultados de la Regresión Logística y de Arboles de Clasificación permite concluir que las variables sexo, estado civil, nivel de educación y la condición de jefe de hogar inciden fuertemente en la probabilidad de que una persona sea desempleada. Se espera que los resultados obtenidos del estudio comparativo, a través métodos de clasificación, sean de gran utilidad para los investigadores de la Economía Laboral, habida cuenta que el desempleo es uno de los problemas que más afecta a la sociedad paraguaya.

Palabras clave: *Desempleo, Modelo Logit, Árboles de Clasificación.*

Abstract: The problem of classification of objects into groups or known populations are great interest in statistics, for this reason, various techniques have been developed to fulfill this purpose. This work tends to identify the risk factors that influence on job insecurity of Paraguayan population; it was adopted on the basis of the Permanent Household Survey 2011. The effect of the predictive variables (age, sex, marital status education level, relationship to head of household, apartment, area, branch and last job category) on the respondent labor situation was estimated through Logistic Regression Analysis and Classification Trees. The analysis of the logistic regression and classification trees results conclude that the variables of sex, marital status, education level and household head status strongly influence on person unemployed probability. It is expected that the results of the comparative study through classification methods are of great value to researchers of Labor Economics, considering that unemployment is one of the problems that most affect Paraguayan society.

Key words: *Unemployment, Logit Model, Classification Trees.*

Introducción

El desempleo es un fenómeno social que afecta de manera negativa, a nivel mundial, en diferentes aspectos de la vida diaria. Paraguay, no es la excepción. Este es el motivo por el que surge la necesidad de estudiar con profundidad esta problemática y las causas que la provocan.

En el año 2008, cerca del 80% de la población paraguaya total se encontraba en edad de trabajar y por ende, medir la dinámica laboral significativa, monitorear la estructura del mercado de trabajo, profundizar sobre el perfil de ocupados y desocupados y ampliar el marco de perspectivas posibles para diagnosticar sobre variados aspectos de la economía y de la sociedad toda, debería ser objeto prioritario de las políticas públicas del país.

Por lo expresado, el análisis del mercado laboral y el estudio de la oferta de trabajo plantean requerimientos metodológicos que son específicamente inherentes a las fuentes estadísticas disponibles.

Recibido: 12/09/2013 Aceptado: 17/10/2022

A los fines de la modelización, se aplican métodos de clasificación supervisada, especialmente indicados cuando las variables predictoras son categóricas, entre los que se encuentran la Regresión Logística y los Árboles de Clasificación (Beltrán, 2011).

En este trabajo se busca identificar los factores de riesgo que inciden en la precariedad laboral de la población paraguaya utilizando como base de información la Encuesta Permanente de Hogares 2011. Para ello, se espera que el estudio comparativo de los distintos métodos de clasificación sea de utilidad a los investigadores de la Economía Laboral, habida cuenta que el desempleo es uno de los problemas que más afecta a la sociedad paraguaya.

La aplicación de diferentes técnicas a un mismo problema tiene por objetivo analizar la eficacia relativa de los diversos instrumentos del análisis multivariado, para brindar la mejor solución al mismo.

Analizar la situación del empleo en el país y de esta manera tener una visión global y crítica del mercado de trabajo, que sirva de base para el diseño e implementación de políticas públicas tendientes a mejorar las condiciones de vida de la población paraguaya es el propósito fundamental de esta investigación.

Estos temas surgen con mayor preponderancia en la actualidad y a partir de la crisis financiera mundial, que puso en relieve el esfuerzo de gran parte de los países para adoptar medidas tendientes a proteger a sus ciudadanos de los efectos negativos, principalmente, en lo referente al empleo.

La Encuesta Permanente de Hogares (EPH) recoge permanentemente datos relacionados con el mercado laboral que no han sido analizados con profundidad, por lo que se pretende aprovechar esta información para conocer los factores más determinantes del desempleo. Para tal fin, se considera útil la aplicación de algunas técnicas de clasificación como la Regresión Logística y Árboles de Clasificación. Estas técnicas se pueden encontrar en Hair et al. (1999), Hosmer & Lemeshow (2000) y Pérez (2006).

Con respecto a los antecedentes en el Paraguay no existen trabajos de investigación relacionados con la comparación de los distintos métodos de clasificación para determinar los principales indicadores que caracterizan el mercado laboral, no obstante, de entre los principales trabajos revisados, de otros países, como en Argentina el trabajo de Díaz *et al*. (2005), denominado Análisis del Desempleo Urbano a Través de un Estudio Comparativo de Métodos de Clasificación.

En este trabajo se identifica los factores de riesgo que inciden en la precariedad laboral de la Población Económicamente Activa. Adopta 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 estima 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 aplican los métodos de Redes Neuronales y Vecino más Cercano.

En Colombia, hay varias investigaciones que guardan relación con la comparación de métodos de clasificación, entre los cuales podemos mencionar la tesis doctoral titulada Comparación de Árboles de Regresión y Clasificación y Regresión Logística (Serna Pineda, 2009) y el trabajo Comparación entre tres Técnicas de Clasificación (Hernández & Correa, 2009).

En el primer artículo mencionado se presenta la comparación, mediante simulación Monte Carlo, de dos técnicas estadísticas de clasificación: Árboles de Regresión y Clasificación (CART) y Regresión Logística. El comportamiento de las técnicas fue medido con la Tasa de Mala Clasificación (TMC). Se presenta una aplicación a la Encuesta de Innovación y Desarrollo Tecnológico, utilizando las técnicas estudiadas, para contribuir a un mejor conocimiento del sistema nacional de innovación en Colombia, donde la Regresión Logística presenta una Tasa de Mala Clasificación más baja que los Árboles de Clasificación.

El segundo artículo muestra los resultados de un estudio de comparación mediante simulación de

tres técnicas de clasificación, Regresión Logística Multinomial (MLR), Análisis Discriminante No métrico (NDA) y Análisis Discriminante Lineal (LDA). El desempeño de las técnicas se midió usando la tasa de clasificación errónea. Se encontró que las técnicas MLR y LDA tuvieron un desempeño similar y muy superior a NDA cuando la distribución multivariada de las poblaciones es normal o logit-normal.

Materiales y métodos

A los efectos de determinar los factores que inciden en el desempleo de la población paraguaya se adoptó como base de información los datos recopilados por la Encuesta Permanente de Hogares (EPH) del año 2011, cuyo levantamiento se realizó en el periodo de octubre a diciembre del mencionado año.

La EPH tiene propósitos múltiples y releva información sobre hogares y personas en torno a las siguientes temáticas: situación laboral, características demográficas básicas (edad, sexo, etc.), características migratorias, habitacionales, educacionales e ingresos. Constituye una de las principales fuentes de información estadística del

Paraguay, que contiene valiosa información sobre las características del empleo y el acceso a beneficios laborales y sociales para una muestra representativa. Cubre todos los Departamentos del país, excluyendo a los Departamentos de Alto Paraguay y Boquerón, cuyas poblaciones representan menos del 2 % de la población total del país. La ejecución de la Encuesta Permanente de Hogares 2011 implico la entrevista de 19,740 personas correspondientes a 4,894 hogares de las áreas urbanas y rurales del país.

La investigación va dirigida a la población que reside habitual o permanentemente en viviendas particulares. Esta población la constituyen las personas en edad de trabajar y comprende a todos los individuos de 10 años y más de edad que suministran mano de obra disponible para la realización de una actividad económica. Se excluye de la investigación a la población residente en viviendas colectivas. El periodo de referencia para captar los datos sobre empleo y desempleo son los últimos 7 días respecto a la fecha de la entrevista.

Las definiciones utilizadas en la EPH están basadas en las recomendaciones de la XIII Conferencia Internacional de Estadísticas del Trabajo, realizada por la Organización Internacional del Trabajo (OIT), en Ginebra en 1982 (DGEEC, 2011). La presente investigación es cuantitativa porque se clasifican los datos de acuerdo a la presencia o ausencia de cierta característica. Es descriptiva por que estudia las relaciones entre cada una de las variables explicativas y la variable dependiente, y es también un análisis multidimensional pues se aplica la técnica de la Regresión logística para estudiar la asociación entre las variables independientes con relación a la variable dependiente. El diseño de investigación es del tipo no experimental porque no existe manipulación de las variables.

La preparación de la base datos consiste en recodificar y agrupar las variables de acuerdo a las categorías de presencia o ausencia del evento, en el análisis exploratorio se obtiene información acerca del comportamiento de cada una de las variables de interés, como de su relación con la variable dependiente, luego se aplica las técnicas de Regresión Logística y Arboles de Clasificación, por último, la verificación del ajuste y validez de los modelos seleccionados.

Resultados y discusión

Un análisis descriptivo de la Situación Laboral en función de las variables que, a priori, se consideran importantes para explicar la condición de empleado o desempleado de un individuo, se realiza en esta sección.

En primer lugar, se eliminan los casos en que se presentan celdas vacías, por lo que, el tamaño muestral queda reducido a 9536 individuos.

Se puede apreciar que el 46,81% de los individuos encuestados en Asunción y Áreas, tanto urbanas como rurales, de la Región Oriental del Paraguay, son desempleados, mientras que el 53; 2% están empleados.

Puntualmente pueden destacarse los siguientes detalles relacionados con las variables que se proponen para estudiar el desempleo:

Los porcentajes menores de desempleados se dan en los Departamentos de San Pedro e Itapúa (8,6% y 7,5 %, respectivamente). La asociación o dependencia entre esta variable y la Situación Laboral es significativa ($p = 9{,}061\ e^{-7}$).

En el Área de Residencia se observa que el porcentaje de desempleados es mayor para aquellos que residen en áreas urbanas. Sin embargo, esta característica no muestra dependencia con la Situación Laboral ($p = 0{,}2308$).

Los Grupos de Edades exhiben porcentaje similares tanto de empleados como de desempleados. De acuerdo con el valor del estadístico chi-cuadrado esta variable tampoco está relacionada con la Situación Laboral ($x^2 = 19747$; $p = 0{,}5777$).

En cuanto al Sexo, el 55; 7% de los desempleados son mujeres y la dependencia entre las dos variables es significativa ($x^2 = 534{,}0344$; $p = 2{,}2\ e^{-16}$).

De manera llamativa se observa que aquellas personas que se encuentran con pareja y son jefes de hogar presentan menores porcentajes de desempleados (31% y 17,2% respectivamente). Esto puede deberse a que personas que están con parejas y sustentan la casa tienen más obligación de tener un empleo formal. Las variables Estado Civil y Jefe de Hogar están relacionados con la Situación Laboral.

La Cantidad de Personas en el Hogar presenta porcentajes similares con relación a los empleados y desempleados, siendo la asociación o dependencia con la Situación Laboral significativa según la prueba x^2($p = 5{,}858\ e^{-13}$).

Teniendo en cuenta la Rama del Último Empleo, se observa que los empleados del Sector Terciario son los menos afectados por el desempleo, ya que el 54% de este sector tienen trabajo. Teniendo en cuenta el estadístico chi-cuadrado se observa que la Rama de Actividad está relacionada con la Situación Laboral ($x^2 = 19.5146$; $p = 5{,}787\ e^{-05}$).

Por lo que a Categoría del Último Empleo se refiere, ésta está relacionada con la Situación Laboral solo al 5% ($p = 0{,}018$). Se puede apreciar que las personas que trabajaban en relación de dependencia, son los más afectados por la falta de trabajo, ya que el 54,7% están desempleadas, así también, este problema afecta a los trabajadores independientes donde el 45,3% de ellos se encuentran sin trabajo.

Finalmente, relacionada con la educación, la categoría Nivel Básico es la que presenta mayor porcentaje de desempleados. Una hipótesis es que esto se debe a que la mayoría de las personas de este nivel realizan tareas menos calificadas. La relación entre esta variable y la Situación Laboral es nuevamente significativa ($p = 2{,}2\ e^{-16}$).

Por lo tanto existen diferencias porcentuales entre empleados y desempleados con respecto a cada una de las variables analizadas, siendo algunas explicativas más relevantes que otras. La "Edad" y "Área de Residencia" del entrevistado fueron las únicas variables no significativas, por lo que se decidió no incluirlas en la propuesta de modelos a ajustar por los métodos de clasificación.

Luego, el ajuste de modelos a través de Regresión Logística y Árboles de Clasificación se realiza con las ocho (8) variables que de acuerdo con la prueba x^2 del análisis previo dependen y por ende, explican la condición de empleado/desempleado.

Variables indicadoras del desempleo a través del Modelo Logit

Para seleccionar las variables que definen el mejor modelo se utiliza el procedimiento "Stepwise". En primer lugar se ajusta el modelo sin variables y a continuación se irán añadiendo y/o eliminando variables en cada paso.

En cada paso se realiza un Test Condicional de Razón de Verosimilitudes para contrastar el modelo del paso anterior con cada uno de los posibles modelos planteados en el nuevo paso. En base a él, se decide qué variable debe entrar (o salir) en ese paso, si procede.

Siguiendo el proceso se llega al modelo final, donde la variable de interés (condición de desempleo) puede ser explicada por el modelo que incluye las variables Jefe de Hogar, Nivel de Educación, Estado Civil, Sexo, Rama del Último Empleo, Departamentos y Categoría del Último Empleo.

El modelo final queda resumido en la Tabla 1, donde se han incluido los coeficientes β_i, las exponenciales de dichos coeficientes (OR), intervalo de confianza para las OR al 95% y el p correspondiente a cada variable o categoría obtenido a través

Tabla 1. Estimación de los coeficientes de ventaja para el modelo final.

Variable	B	OR	IC 95% OR	p-valor
Constante	-1.53	0.22	(0.17; 0.27)	< 2e-16
Jefe de Hogar				
No Jefe/a*				
Jefe/a	0.81	2.24	(2.00; 2.50)	< 2e-16
Nivel de Educación				
Nivel Básica *				
Nivel Media	-0.62	0.54	(0.49; 0.59)	< 2e-16
Nivel Superior	-1.50	0.23	(0.19; 0.27)	< 2e-16
Estado Civil				
Con Pareja*				
Sin Pareja	1.04	2.84	(2.57; 3.14)	< 2e-16
Sexo				
Hombres*				
Mujeres	1.21	3.34	(3.01; 3.70)	< 2e-16
Rama del Ultimo Empleo				
Sector Primario*				
Sector Secundario	0.14	1.15	(0.98; 1.35)	0.093
Sector Terciario	0.492	1.635	(1.43; 1.87)	1.0 e^{-12}
Departamento				
Asunción*				
San Pedro	-0.27	0.77	(0.62; 0.95)	0.013
Caaguazú	0.02	1.03	(0.85; 1.24)	0.797
Itapúa	-0.42	0.66	(0.53; 0.81)	8.9 e^{-05}
Alto Paraná	-0.29	0.75	(0.63; 0.90)	0.001
Central	-0.04	0.97	(0.82; 1.14)	0.669
Resto	0.02	1.02	(0.86; 1.22)	0.787
Categoría del Último Empleo				
Trabajador en Relación de Dependencia*				
Trabajador Independiente	-0.22	0.80	(0.719; 0.88)	2.9 e^{-05}
*Categoría de Referencia				

del test de Wald.

Para el ajuste global del modelo se calculan la Tasa de Clasificaciones Correctas y el Área bajo la Curva ROC.

Al calcular la Tasa de Clasificaciones Correctas. Se calcula la función para los puntos de corte entre 0,1 y 0,9, (de 0,1 en 0,1), para encontrar cuál de ellos maximiza la Tasa de Clasificaciones Correctas. El punto de corte que maximiza la Tasa de Clasificaciones Correctas se encuentra alrededor de 0,5. Para conseguir estimaciones más precisos se prueban los puntos de corte entre 0,45 y 0,54 (de 0,01 en 0,01). La conclusión es que el mejor punto de corte es 0,51, que da como resultado una Tasa de Clasificaciones Correctas de 69,91401 (≈ 70 %), y por ende, la capacidad predictiva del modelo es relativamente buena.

El área bajo la curva ROC representa la probabilidad de que un individuo desempleado tenga un valor en la escala de medida considerada mayor que un individuo con probabilidad de no estar desempleado.

El valor del estadístico, equivalente al área bajo la curva ROC, es de 0,761, que se considera aceptable para evaluar el modelo propuesto.

El Análisis de los Residuos y la Distancia de Cook (medidas de influencia) son los indicadores aplicados para realizar la Validación y Diagnosis del modelo.

El resumen del Análisis de los residuos se tienen como valor mínimo -2,0799 y como valor máximo 2,7178, lo que en principio indica que podría haber valores que distorsionan el ajuste global del modelo. Sin embargo, también se tiene que q_1 = -0,9725 y q_3= 0,9789 o sea que el 50% de los residuos se encuentran entre ±1, por lo que es de esperar que entre ±2 se encuentre la mayoría de los datos (≈ el 95% si estos fueran normales).

Identificando cada uno de los residuos de los individuos encuestados, de 9425 residuos, 111 son mayores a 2 (o sea, apenas un 1% está fuera del rango esperado). No obstante, a pesar de que esta situación no es alarmante, las medidas de influencia, específicamente las Distancias de Cook, son una prueba útil para la identificación de puntos

influyentes.

El valor máximo de las distancias de Cook es 0,001135235, con lo que se concluye finalmente que no hay ninguna distancia mayor que 1 y por tanto, se acepta el exceso detectado en los residuos con lo que, el modelo queda validado.

Variables indicadoras del desempleo a través de Árboles de Clasificación

Para realizar una partición de los datos en subconjuntos que sean homogéneos con respecto a la variable criterio (Empleado/Desempleado), se utilizan las ocho variables que mostraron estar asociadas a la "Situación Laboral" (Departamento, Jefe de Hogar, Sexo, Estado Civil, Número de Personas en el Hogar, Categoría del Último Empleo, Rama del Último Empleo y Nivel de Educación). Aquí tampoco se analizan las variables Área y Grupo de Edades, debido a que en el análisis descriptivo resultaron no estar asociadas a la Situación Laboral (Empleado/ Desempleado).

El modelo resultante presenta las variables y categorías de las mismas, que son discriminantes en el análisis de caracterización de los desempleados. El diagrama de árbol ofrece en el nodo raíz la variable dependiente Situación Laboral con el 53,2% de empleados y el 46,8% de desempleados.

La primera división del árbol se realiza a partir de la variable Jefe de Hogar, de tal manera que la categoría "no jefe/a de hogar" aparece con mayores probabilidades de desocupación (56,2 %).

Asimismo, los jefes de hogar se segmentan según el Sexo, y se observa que las mujeres exhiben mayores posibilidades de aparecer desempleadas (51,2 %). Inversamente, los hombres que son jefes de hogar muestran un elevadísimo porcentaje de empleados, que llega al valor del 82,5 %.

Por otro lado, los no jefe/a se clasifican según el Nivel de Educación, y los que ostentan niveles de educación bajos o medios tienen mayores probabilidades de estar desempleados (o en sentido inverso los que poseen una formación superior tienen más probabilidad de estar empleados). Seguramente, lo afirmado se debe a que personas con niveles educativos bajos o medios se desempeñan en tareas menos calificadas y los otros en los niveles de dirección o profesional.

Para los niveles de educación bajos y medios la clasificación se da según el Estado Civil destacándose que las personas que están sin pareja presentan mayores porcentajes de desempleados (73,7% y 58,1 %, respectivamente), lo cual permite confirmar que las personas que se encuentran en pareja realizan esfuerzos mayores para lograr una ocupación. Por otra parte, los que poseen Nivel de Educación Superior se clasifican según el sexo, y como era de esperarse tanto los hombres como las

Tabla 2. Ganancia en el Análisis de la Situación Laboral, según el Árbol de Clasificación obtenido.

	Nodo		Ganancia			
Nodo	N	%	N	%	Respuesta	Índice
8	2045	21.40%	1507	33.80%	73.70%	157.40%
10	1964	20.60%	1142	25.60%	58.10%	124.20%
16	491	5.10%	276	6.20%	56.20%	120.10%
9	1140	12.00%	558	12.50%	48.90%	104.60%
17	257	2.70%	107	2.40%	41.60%	88.90%
11	717	7.50%	281	6.30%	39.20%	83.70%
13	416	4.40%	142	3.20%	34.10%	72.90%
14	1227	12.90%	268	6.00%	21.80%	46.70%
12	296	3.10%	64	1.40%	21.60%	46.20%
15	983	10.30%	119	2.70%	12.10%	25.90%

mujeres exhiben mayores posibilidades de aparecer ocupados (78,4% y 65,9 %, respectivamente).

Analizando la rama del árbol que corresponde a los No Jefe/a, la condición de empleados o desempleados depende del sexo. En tal sentido, las mujeres y los hombres se clasifican nuevamente según su nivel de educación destacándose que en casi todos los niveles los hombres tienen más posibilidades de estar ocupadas que las mujeres.

En resumen, se detectaron 10 clases en el análisis, ordenadas por riesgo decreciente de desempleo, como se observa en la Tabla 2.

En la Tabla 3 se observan las clases (caminos de nodos) ordenados según la proporción de desempleados, es decir, permite saber qué subgrupo de personas tiene mayores probabilidades de ser clasificado como tal.

Se advierte (Tabla 2) que cuando el índice presentado en la última columna supera el 100 %, tal como ocurre en las clases 1, 2, 3 y 4 la proporción de desempleados es superior en esas categorías. Es decir, que las clases No jefe/a con Nivel de Educación Básico y que están Sin Pareja (Índice 157,4), No jefe/a con Nivel de Educación Medio y que están Sin Pareja (Índice 124,2), Jefe/a que son mujeres y tienen un Nivel de Educación Básico (Índice 120,1) y No jefe/a con Nivel de Educación Básico y que están Con Pareja (Índice 104,6) representan los grupos de mayor riesgo de desempleo.

Tabla 3. Clase en el Análisis Empleado-Desempleado.

Clase	Análisis Empleado-Desempleado
1	No Jefes con Nivel de Educación Básico y Sin Pareja
2	No Jefes con Nivel de Educación Medio y Sin Pareja
3	Jefes Mujer con Nivel de Educación Básico
4	No Jefes con Nivel de Educación Básico y Con Pareja
5	Jefes Mujer con Nivel de Educación Medio o Superior
6	No Jefes con Nivel de Educación Medio y Con Pareja
7	No Jefes con Nivel de Superior siendo Mujer
8	Jefes Hombre con Nivel de Educación Básico
9	No jefes con Nivel de Educación Superior siendo Hombre
10	Jefes Hombre con Nivel de Educación Medio o Superior

Tabla 4. Clase en el Análisis Empleado-Desempleado.

	Pronosticado		
Observado	Empleado	Desempleado	% correcto
Empleado	3497	1575	68.90%
Desempleado	1539	2925	65.50%
% global	52.80%	47.20%	67.30%

En el otro extremo, los Jefes que son Hombres y tienen un Nivel de Estudio Medio o Superior, que componen la clase 10, corresponden al grupo de menor riesgo, con un porcentaje de desempleo del 12,1 %, que representa sólo la décima parte del total de la muestra.

Para analizar, las variables de predicción del modelo aparecen en las Tablas de Riesgo y de Clasificación, y proporcionan una rápida evaluación de la bondad del funcionamiento del modelo.

En la Tabla 4 se debe considerar que los resultados brindados por el árbol obtenido son coherentes, en el sentido que el modelo clasifica de forma correcta, aproximadamente, al 67,3% de los individuos.

Conclusión

Utilizando la base de datos de la Encuesta Permanente de Hogares del año 2011 y teniendo en cuenta características demográficas, sociodemográficas y económicas de los trabajadores residentes en Asunción y las áreas, tanto urbanas como rurales, de la Región Oriental del Paraguay, se ha realizado una inferencia sobre la situación de empleado o desempleado de los paraguayos, a través de dos técnicas diferentes, la Regresión Lo-

gística y los Arboles de Clasificación, presentando un modelo matemático caracterizado por factores que influyen para que en esta sociedad se origine el desempleo.

Partiendo de un conjunto de características consideradas predictores del desempleo, "Departamento de Residencia", "Área (urbana o rural) de Residencia", "Edad", ´

"Condición de Jefe de Hogar", "Sexo", "Estado Civil", "Numero de Personas en el Hogar", "Categoría del Ultimo Empleo", "Rama del ´ Ultimo Empleo" y "Nivel de Educación", un análisis inicial exploratorio, pero a su vez confirmatorio, establece que la "Edad" y el "Área de Residencia" ´ no son significativas a la hora de explicar el desempleo de los paraguayos, y por ende, son excluidas de los procedimientos de selección de modelos.

Al recurrir a la Regresión Logística, se demuestra que las variables "Jefe de Hogar", "Nivel de Educación", "Estado Civil", "Sexo", "Rama del Ultimo Empleo",

"Departamentos" y "Categoría del Ultimo Empleo" ´ contribuyen a la construcción de un modelo que permite estimar la probabilidad de que un individuo de nacionalidad paraguaya sea desempleado. La capacidad predictiva del modelo es buena, porque realiza un 70 % de clasificaciones correctas. Otros mecanismos de diagnosis (por ejemplo, área bajo la curva ROC, distancias de Cook, análisis de residuos) también permiten considerar que el modelo propuesto es aceptable.

Sin embargo, la técnica de Arboles de Clasificación arroja un modelo mucho más simple para explicar el carácter de desempleado de un individuo. Este modelo

tiene en cuenta tan solo las variables ´ "Jefe de Hogar", "Nivel de Educación" ´ , "Estado Civil" y "Sexo", las cuales también están involucradas en el modelo logístico obtenido, considerándose este hecho como razón suficiente para tenerlas en cuenta al diseñar políticas tendientes a disminuir los niveles de desempleo.

Al comparar las dos técnicas de clasificación se encontró que el modelo obtenido por Arboles de Clasificación (que posee menos variables y todas relacionadas con condiciones que pueden considerarse "propias" de cada trabajador) presenta una Tasa de Clasificaciones Correctas más baja que la Regresión Logística (que incluye además de las personales, otras características más relacionadas con el tipo y lugar del empleo). ´

También puede mencionarse como ventaja de la Técnica de Arboles el hecho de haber detectado 10 clases (o camino único entre el nodo raíz y cada uno de los nodos terminales del Árbol) en el análisis, ordenadas por riesgo decreciente de desempleo.

Precisamente, esta es la mayor contribución de esta metodología, porque en la lectura de las diferentes ramas correspondientes a esos nodos brinda la caracterización de los perfiles de los casos estudiados en este trabajo, es decir, de los desempleados paraguayos.

La aplicación de métodos de clasificación permite obtener más y mejor información sobre los indicadores del desempleo (sexo, estado civil, nivel de educación y la condición del jefe de hogar), para el diagnóstico y actualización de la situación de los pobladores de la región oriental de Paraguay.

Se considera que esta investigación permite comenzar a identificar indicadores del desempleo que afecta a los pobladores de la Región Oriental del Paraguay. Es un paso inicial que puede encaminar estudios que permitan a establecer políticas tendientes a ampliar el marco de perspectivas laborales posibles, así como diagnosticar otros aspectos de la economía y de la sociedad paraguaya.

Conflictos de interés: Los autores declaran no tener conflictos de interés.

Contribución de los autores: Los autores contribuyeron manera equitativa en la elaboración de este artículo.

Referencias

Beltrán, C. (2011). Análisis de Regresión Logística Aplicado a la Clasificación Textos Académicos: Biometría y Filosofía. *Revista de Epistemología y Ciencias Humanas*, 3: 50–60.

DGEEC (Direccion General y Estadísticas y Censos). (2011). *Principales Resultados de la EPH 2011*. [Consulted: 12.viii.202]. <http://www.dgeec.gov.py>. Asunción: Paraguay. 33 pp.

Díaz, M., Ferrero, F., Díaz, C., Caro, P & Stimolo, M.I. (2005). Análisis del Desempleo Urbano a Través de un Estudio Comparativo de Métodos de Clasificación. *Revista de Economía y Estadística*, 43(2): 61–83.

Hair, J.F.Jr., Anderson, R.E., Tatham, R.L. & Black, W.C (1999). *Análisis Multivariante*. (5ta. Edición). Hoboken: Prentice Hall. 799 pp.

Hernández, F. & Correa, J.C. (2009). Comparación entre tres Técnicas de Clasificación. *Revista Colombiana de Estadística*, 32(2): 247–265.

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*. (2° Edición). New York: Wiley-Interscience. 375 pp.

Pérez, J.M. (2006). *Árboles Consolidados: Construcción de un Árbol de Clasificación basado en múltiples submuestras sin renunciar a la explicación*. (Tesis doctoral). Donostia: Universidad del País Vasco. 271 pp.

Serna Pineda, S. C. (2009). *Comparación de árboles de regresión y clasificación y regresión logística*. (Disertación de maestría). Medellín: Universidad Nacional de Colombia. 60 pp.