Problem Set 3: Predicting Poverty

Presentado por Stiven Peralta, Jazmine Galdos, Andrea Clavijo, Sergio Jiménez y Nicolás Barragán

I. Introducción

Este trabajo busca construir un modelo predictivo que permita clasificar los hogares en condición de pobreza a partir de la información recolectada por el DANE, en el marco de la Misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad (MESEP)[1]. El interés principal de este estudio surge de la necesidad de identificar aquellos hogares que se encuentran por debajo de la línea de pobreza para enfocar las políticas públicas que buscan reducir esta problemática, tales como la focalización y distribución de subsidios a poblaciones vulnerables.
Para la definición del modelo predictivo se construye una base que consolida datos relevantes a nivel de hogar, considerando características del jefe del hogar (sexo, edad, nivel educativo, tipo de trabajo, entre otras), del grupo familiar (número de integrantes, porcentaje de ocupación, promedio de años de ocupación, entre otras), variables de ingreso (ingresos laborales, primas, subsidios, entre otras) y características de la vivienda (número de habitaciones, condición de tenencia, entre otras).
Para identificar el modelo predictivo con mejor ajuste se definieron dos aproximaciones, por un lado, se plantearon modelos de clasificación tomando como variable independiente la dummy que indica si un hogar este clasificado como pobre (1) o no (0). Por otro lado, se plantearon modelos de predicción del ingreso de los hogares, para posteriormente identificar si el hogar es pobre o no con respecto a la línea de pobreza. Como resultado, se identifica que ambas aproximaciones resultan en una precisión (Accuracy) entre 78% y 86%, donde el modelo que presenta el mejor ajuste corresponde al modelo de clasificación a través de regresión logística con regularización por elastic-net (modelo 4) con una precisión del 86%.
Como principal conclusión, no se evidencian mayores diferencias en la capacidad predictiva de los diferentes modelos desarrollados, por lo que las mejoras de ajuste se dan por la inclusión y/o exclusión de variables de interés, en donde las principales variables para la predicción son el porcentaje de ocupados, la edad, el tipo de trabajo e interacciones con el género del jefe del hogar. Finalmente, resulta relevante para la selección del modelo la menor cantidad de predictores y el menor costo computacional requerido para su estimación.

Nota: *la base de datos usada, al igual que los scripts de R y el presente documento están disponibles en el repositorio de GitHub en el siguiente enlace: https://github.com/stivenperalta/Problem_set_3*

II. Antecedentes

Combatir la pobreza es uno de los desafíos más relevantes de los estados e instituciones en la actualidad. Esto, dado su impacto en la calidad de vida de millones de personas alrededor del mundo. La pobreza no solo afecta la capacidad de satisfacer las necesidades básicas (vivienda,

[1] *La Misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad MESEP fue creada en enero de 2009 mediante un convenio DANEDNP con el objetivo de evaluar los factores que afectaron la comparabilidad de las cifras de mercado laboral y pobreza con el paso de la Encuesta Continua de Hogares ECH a la Gran Encuesta Integrada de Hogares GEIH, y realizar los empalmes correspondientes. A partir de la MESEP se adopta la nueva metodología para la medición de pobreza monetaria en Colombia, que adopta cambios tanto en la línea de pobreza como en la construcción del agregado de ingreso del hogar. Esta metodología se acerca más a las metodologías vigentes en los demás países latinoamericanos, con lo cual se facilita la comparabilidad en el contexto regional.* (DANE - DNP, 2012)

salud, alimentación), sino que también limita el acceso a oportunidades y perpetua los ciclos de pobreza generacional por medio de la desigualdad social (World Bank, 2018).

A lo largo del tiempo se han implementado diferentes aproximaciones a esta problemática, las cuales incluyen el desarrollo de políticas públicas y programas de asistencia para atender las necesidades básicas de la población más vulnerable. No obstante, uno de los retos más relevantes es contar con una medición adecuada que permita un direccionamiento eficiente de los esfuerzos para atenderla; especialmente, al tener en consideración los recursos limitados para afrontar este fenómeno. La medición de la pobreza se ha convertido en uno de los temas de estudio frecuentes en la econometría (Kumar, Gore, & Sitaraman, 1996) pasando por aproximaciones diversas que consideran factores cuantitativos como definición de rangos mínimos de ingreso, aproximaciones subjetivas relacionadas con la percepción y/o capacidad de acceder a un mínimo de bienes o servicios, así como aproximaciones multidimensionales que buscan considerar diversos factores para la definición de la población en condición de pobreza (Nunes, 2008).

Para el caso colombiano, se pueden identificar dos métodos para la medición de la pobreza. Por un lado, un método directo que evalúa a capacidad de los hogares para satisfacer una serie de necesidades identificadas, teniendo como ejemplo los indicadores de necesidades básicas insatisfechas (NBI) y el índice de pobreza multidimensional (IPM). Así mismo, también se puede identificar una metodología indirecta, que es la planteada por la MESEP, para categorizar los hogares en condición de pobreza a partir de la definición de rangos de ingreso (DANE - DNP, 2012). Sin embargo, resulta complejo contar con información primaria que permita la medición de los niveles de ingreso y clasificación de los hogares de manera constante para optimizar la implementación de las políticas públicas que buscan hacer frente a la pobreza, por lo que se pueden generar ineficiencias en el direccionamiento de los esfuerzos (World Bank, 2018).

Es en este contexto, es de gran relevancia contar con modelos de machine learning que, a partir del análisis de datos, pueden identificar patrones y tendencias que permiten predecir las condiciones que hacen a un hogar propenso a estar en condición de pobreza (Zixi, 2021). Estos modelos permiten hacer una identificación más rápida y menos costosa para la adecuada implementación de las políticas públicas. Al enfocar los recursos en los lugares donde se necesita con mayor urgencia, los gobiernos y las organizaciones pueden maximizar el impacto de sus intervenciones y optimizar el uso de los recursos disponibles (World Bank, 2018). Los modelos predictivos también pueden ayudar a identificar a los individuos y familias que requieren asistencia de manera más inmediata, lo cual permite una respuesta rápida y efectiva.

III. Datos

Para el desarrollo del modelo predictivo, se tomaron los datos de la *Misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad* (MESEP) del Departamento Administrativo Nacional de Estadística (DANE, 2018), los cuales toman los datos de la *Gran encuesta integrada de hogares* (GEIH)[2]. Esta base de datos contiene información sobre características

[2] La GEIH recolecta información acerca de las condiciones de empleo vigentes en el país: trabajo, tipo de trabajo, salarios, ingresos, entre otros; de igual forma, recolecta variables demográficas de los encuestados como el sexo, la edad, el nivel educativo. De acuerdo con el DANE, el universo de esta encuesta está conformado por la población civil no institucional, residente en todo el territorio nacional.

de individuos y unidades de gasto, junto con el valor de la línea de pobreza e indigencia. Los datos se descargaron en la página web Kaggle[3].

Para este ejercicio, hay 4 bases de datos divididas en *train* (2 bases: una de hogares y otra de personas) y *test* (2 bases: una de hogares y otra de personas). La limpieza de los datos comenzó por identificar qué variables se encontraban en la base *test* de personas y no estaban en la base *test* de hogares; posteriormente, se procedió a agregar estas variables a la base por hogar (mediante un *merge* por id del hogar). Este ejercicio se realizó con el objetivo de maximizar el número de variables correlacionadas con la variable de interés (pobreza). Luego, se comparó qué variables existían tanto en la base *train* de hogares como en la base *test* de hogares, para asegurarse de tener las mismas variables y así poder hacer la predicción de los modelos. El *merge* de personas a hogares se hizo para poder crear variables que permitieran tener información agrupada por hogar. Esto, además, permitió crear variables adicionales de hacinamiento (número de personas por cuarto), porcentaje de ocupados por hogar y grados de educación consolidado por hogar, entre otros.

Para realizar el cruce de información correspondiente entre la base *test* de personas a *test* de hogares se tuvo en cuenta la información del jefe de hogar. Esto ya que la gran mayoría de variables disponibles en la base *test* de personas son categóricas. Por este motivo, tener en cuenta las características a nivel de jefe de hogar, como aquellas que son relevantes para calcular pobreza por hogar, es oportuno al evaluar la pobreza teniendo en cuenta que, según el DANE, el jefe de hogar es la persona que aporta más dinero a este, razón por la que se considera como el principal sostén económico de la familia. Sumado a esto, también se creó otra variable que distinguiera si ese jefe de hogar es mujer para tener en cuenta posibles brechas de género que incidan en la probabilidad de que un hogar sea pobre. Una vez fueron creadas y agregadas las variables por hogar, se hizo el ajuste de los *missing values* haciendo imputación por distinción de ocupados en aquellas variables relacionadas con el sector formal y, en las restantes, al ser en su mayoría variables categóricas, por la moda respectiva en cada una de ellas (para mayor detalle véase Anexo 1).

En total, después de realizada la agregación, limpieza e imputación se cuenta con una base de 231.128 observaciones distribuidas en una muestra de entrenamiento (***train***) con 164.960 observaciones (71%) y otra de prueba (***test***) con 66.168 observaciones (29%)[4]. Al final, el set contiene 48 variables con las siguientes características: 31 variables categóricas, 16 variables continuas y 1 variable de *id* del hogar (véase Anexo 2 para consultar directorio de variables).

La Tabla 1 resume las estadísticas descriptivas de las variables de interés. Se puede observar que en la base ***train*** el promedio del ingreso per cápita por hogar es de COP 870.639, sin embargo, la mediana de esta variable se ubica muy por debajo del salario mínimo del 2018[5], al llegar a COP 543.568. También se encuentra que, en la muestra, en promedio, hay 3,4 cuartos por hogar y 3,3 personas por hogar y la edad media del jefe de hogar es de 49,6 años. Por otra parte, la media de horas trabajadas por semana es 33,21, mientras que la mediana es 40 horas. Este dato llama la atención ya que en Colombia se exigían 48 horas trabajadas a la semana para el año 2018. Las estadísticas descriptivas de variables categóricas utilizadas se encuentran en el Anexo 1. En cuanto a variables asociadas al mercado laboral, la mayor concentración de la distribución se observa en la variable otros, cuyo resultado está en concordancia con lo

[3] https://www.kaggle.com/competitions/uniandes-bdml-202313-ps31
[4] La base inicial contaba con 231.128 observaciones de hogares (train y test) y 762.753 observaciones de personas (train y test).
[5] COP 781.242.

resaltado por el Banco de la República (s.f.), quien sostiene que, en Colombia, el mercado del trabajo es mayoritariamente informal.

De igual forma, al validar el nivel educativo se encuentra una distribución más concentrada en básica primaria, media y universitaria. Respecto a esta última, es interesante ver cómo este nivel educativo evidencia casi la misma distribución con básica primaria, tendiendo está última la mayor participación en toda la muestra train (Gráfico 1).

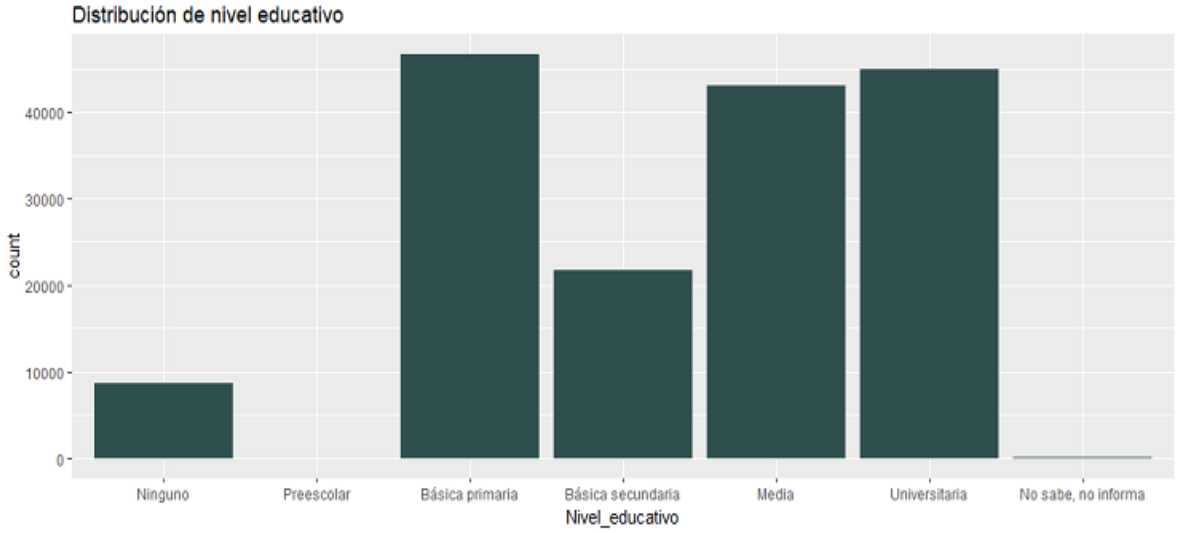
Tabla 1. Estadísticas descriptivas de variables continuas

	Test					**Train**				
	Min.	**Max.**	**Promedio**	**Mediana**	**Desv. Est.**	**Min.**	**Max.**	**Promedio**	**Mediana**	**Desv. Est.**
Ingreso per cápita por hogar	-	-	-	-	-	0	88.833.333	870.639	543.568	1.244.350
Porcentaje ocupados	0	1	0,49	0,5	0,32	0	1	0,49	0,5	0
Cuartos por hogar	1	18	3,4	3	1.20	1	98	3,39	3	1.24
Personas por hogar	1	21	3,3	3	1.80			3,3	3	1.77
Personas por cuarto	0,2	14	1,73	1,5	0.84	0,2	16	1,7	1,5	0.83
Edad Jefe de hogar	11	101	49,69	49	16.37	11	108	49,6	49	16.39
Horas trabajadas por semana	0	130	33,21	40	24.78	0	130	33,32	40	24.89

Adicionalmente, al explorar posibles variables asociadas al género se creó una que hiciera la distinción sobre si el jefe del hogar es mujer o no. Se encontró que en el 42% de los hogares la jefe del hogar es mujer; dicha proporción evidencia una muestra casi balanceada respecto a esta variable. Sumado a esto, también fue posible constatar que el porcentaje más predominante en cuanto a tipo de vivienda es en arriendo o alquiler con un 39%, seguido de vivienda pagada en un 38%[6]. Respecto a las variables asociadas al ingreso, teniendo en cuenta que estas son de tipo categórica, se observa que, en su mayoría los encuestados manifiestan no percibir ingresos adicionales relacionados con actividades laborales diferentes a su principal fuente de ingresos o subsidios de alimentación, primas o bonificaciones.

Gráfico 1. Análisis descriptivo

[6] Para un análisis más detallado, véase Anexo 3.



IV. Modelo y resultados

A. Modelos de clasificación

Inicialmente se plantea una aproximación a partir de modelos de clasificación buscando predecir directamente si un hogar se encuentra por debajo de la línea de pobreza a partir de las diferentes características de los hogares. Se plantean modelos de regresión logística (Logit) con regularización (elastic net), vecinos cercanos (KNN), análisis discriminativo lineal (LDA) y cuadrático (QDA), Bayes ingenuo (Naive Bayes) y arboles de decisión (AdaBoost), los cuales toman la forma:

$$\boldsymbol{Hogar_{pobre}} = \boldsymbol{f(x)}$$

Siendo $\boldsymbol{Hogar_{pobre}}$ una dummy de pobreza que toma el valor de 1 si el hogar está clasificado por debajo de la línea de pobreza y 0 en caso contrario y $\boldsymbol{x}$ un vector que incluye diferentes características de los hogares; entre estos se incluye: sexo, edad, nivel educativo, tipo de trabajo, nivel de ingresos del jefe del hogar, número de personas por hogar, ocupación y educación de los integrantes del hogar, así como características de la vivienda e interacciones respecto al género del jefe del hogar en algunos casos.

a. Regresión logística con regularización (Logit -Elastic Net)

En la Tabla 2 se muestran los resultados generales para los cinco modelos de regresión logística con regularización por elastic net desarrollados dentro de muestra (train), para los cuales se utilizó una grilla de hiperparametros donde $\alpha = \in \{0.855, 0.865\}$ y $\lambda = \in \{0.0000, 0.0005\}$ y se realizó el control de entrenamiento por validación cruzada con $k\ folds = 10$.

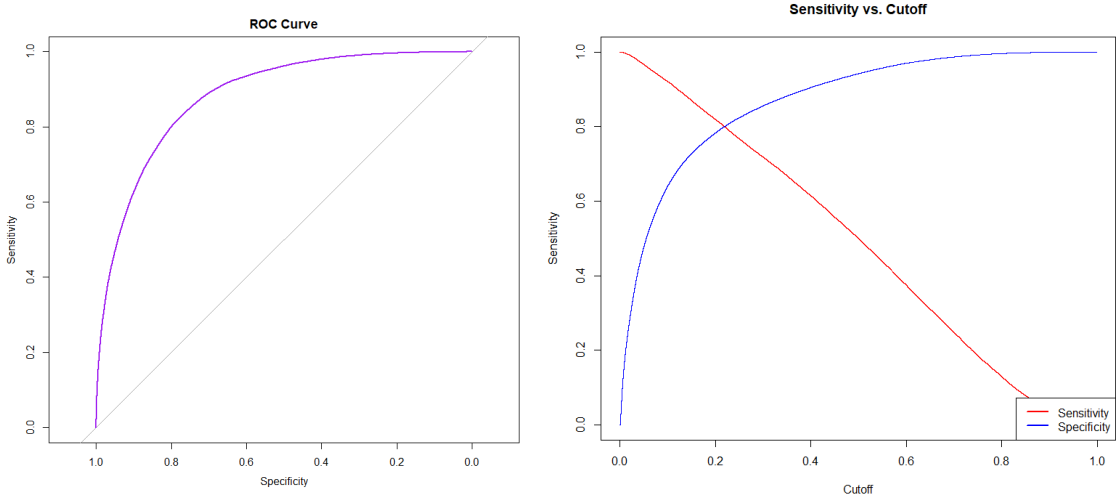
Tabla 2. Resultados Modelos de Clasificación LOGIT

	(1)	(2)	(3)	(4)	(5)
Métricas de Desempeño					
Precisión	**0,8279**	**0,8535**	**0,8391**	**0,8612**	**0,8537**
Sensibilidad	0,9506	0,9433	0,8820	0,9410	0,9501
Especificidad	0,3380	0,4947	0,6676	0,5423	0,4683
Matriz de Confusión					

Verdaderos Positivos (TP)	125.412	124.452	116.367	124.156	125.358
Verdaderos Negativos (TF)	11.162	16.336	22.047	17.908	15.466
Falsos Positivos (FP)	21.862	16.688	10.977	15.116	17.558
Falsos Negativos (FN)	6.524	7.484	15.569	7.780	6.578
Parámetros de Regularización					
alpha (α)	0,8600	0,8650	0,8650	0,8550	0,8550
lambda (λ)	0,0000	0,0010	0,0010	0,0004	0,0001
Ajustes al Modelo					
Corte de clasificación	0,50	0,50	0,35	0,50	0,82
Pliegues de validación cruzada	10	10	10	10	10
Variables correlacionadas	Sí	No	No	No	No
Interacciones según género del jefe de hogar	No	No	No	Sí	Sí
Balanceo de muestra	NA	NA	NA	NA	Up-train

El modelo (1) plantea el primer intento de modelo con una selección de variables de interés que intuitivamente podrían influir en la clasificación de un hogar como pobre. Los siguientes modelos (2-5) se arman agregando nuevas variables (incluyendo interacciones) y descartando aquellas variables que se encuentran más correlacionadas entre sí. La diferencia entre los modelos (2) y (3) corresponde a la variación del corte de clasificación, en donde inicialmente se usó el clasificador de Bayes (0.5), pero dada la alta sensibilidad se planteó que modificar este clasificador podría mejorar el ajuste del modelo como se evidencia en el Gráfico 2. Los resultados indican que al reducir el corte de clasificación se reduce la precisión del modelo en aproximadamente un 2%.

Gráfico 2.ROC - Cut off Modelo (2)



Los modelos (4) y (5) incluyen interacciones entre las variables de interés y el género del jefe del hogar, teniendo en cuenta la evidencia que se tiene de que los hogares con jefatura de mujeres tienen una mayor vulnerabilidad frente a la pobreza (CEPAL, 1990). El modelo (5) plantea el balanceo de la muestra a través de la replicación aleatoria de las observaciones de la clase con menor representación, de manera que la muestra pasa de 164.960 observaciones a

263.872 observaciones (up-train)[7]. Adicionalmente, para este caso se ajusta el corte de clasificación a 0.82, sin embargo, con estos ajustes no se evidencian mejoras en la precisión del modelo.

b. Otros modelos de clasificación (KNN, LDA, QDA, NB, ADABOOST)

En la Tabla 33 se muestran los resultados para los otros seis modelos de clasificación estimados, dentro de muestra (train), los cuales en general no presentan mejores estimaciones que los modelos de regresión logística anteriormente desarrollados.

Tabla 3. Resultados otros modelos de clasificación

		precisión
(6)	KNN	0,8062
(7)	LDA	0,8268
(8)	LDA	0,8505
(9)	QDA	0,7830
(10)	QDA	0,7868
(11)	NB	0,8153

El modelo (6) corresponde a una clasificación por vecinos cercanos (KNN), en donde se identificó que el mejor modelo bajo esta metodología considera 11 vecinos. Cabe resaltar que este modelo supone una complejidad computacional bastante alta, y al no identificar mejoras en la predicción, no se considera un modelo viable para la predicción.
Los modelos (7) y (8) corresponden a modelos de análisis discriminativo lineal (LDA), mientras que los modelos (9) y (10) corresponden a modelos de análisis discriminativo cuadrático (QDA), en donde los modelos (8) y (10) presentan una precisión levemente superior al considerar las interacciones respecto al género del jefe del hogar. En general, los modelos de análisis discriminativo cuadrático suponen una peor predicción que los modelos lineales, siendo el modelo (8) el que presenta una mejor predicción en este grupo. Para estos modelos es fundamental restringir los predictores con una alta correlación o poca varianza.
El modelo (11) corresponde a un Bayes Ingenuo (Naive Bayes) donde se deben reducir considerablemente el número de predictores (pasando de 13 a 8 variables) para ajustar la muestra a la distribución normal que asume este método. Finalmente, se construyó un modelo de árboles de decisión que se descartó por la enorme complejidad computacional que suponía (AdaBoost)[8].

B. Modelos de predicción del ingreso

Por otro lado, se plantean modelos para predecir primero el ingreso de los hogares, y posteriormente clasificar como pobres aquellos hogares cuyo ingreso predicho se encuentre por debajo de la línea de pobreza. Se plantean modelos de regresión con regularización (elastic net), los cuales toman la forma:

$$(1) \qquad \widehat{ing} = f(x)$$

[7] Es importante recalcar que se emplearon técnicas de balanceo de la variable de interés como down-sample; no obstante, las predicciones obtenidas no mejoraron la precisión de predicción (accuracy), en comparación con la base desbalanceada, de los modelos, razón por la que se desistió en seguir empleando esta técnica de balanceo en los demás modelos estimados.
[8]Se estuvo entrenando el modelo, tanto por clasificación como por regresión, y posterior a las 30 horas de entrenamiento, se optó por cancelarlo debido a su alto costo computacional.

$$(2) \quad \boldsymbol{Hogar_{pobre}} = \begin{cases} \mathbf{1}, & \boldsymbol{\widehat{ing} < Lp} \\ \mathbf{0}, & \boldsymbol{\widehat{ing} \geq Lp} \end{cases}$$

Siendo $\boldsymbol{\widehat{ing}}$ el ingreso estimado del hogar y $\boldsymbol{x}$ un vector que incluye diferentes características de los hogares (sexo, edad, nivel educativo y tipo de trabajo del jefe del hogar, número, ocupación y educación de los integrantes del hogar, así como características de la vivienda). $\boldsymbol{Hogar_{pobre}}$ corresponde a una dummy de pobreza que toma el valor de 1 si el ingreso estimado del hogar $\boldsymbol{\widehat{ing}_i}$ es inferior al valor definido para la línea de pobreza $\boldsymbol{Lp_i}$ y 0 en caso contrario.

En la Tabla 4tabla 4 se muestran los resultados generales para los modelos de regresión del ingreso con regularización por elastic net desarrollados, para los cuales se realizó el control de entrenamiento por validación cruzada con $k\ folds = 5$.

Tabla 4.Resultados de modelos de predicción de ingreso

	(12)	**(13)**
Métricas de Desempeño		
Precisión	**0,8387**	**0,9418**
Sensibilidad	0,9426	0,9418
Especificidad	0,4237	na
Matriz de Confusión		
Verdaderos Positivos (TP)	124.360	155.364
Verdaderos Negativos (TN)	13.992	0
Falsos Positivos (FP)	19.032	0
Falsos Negativos (FN)	7.576	9.596
Parámetros Regularización		
alpha (α)	0,9000	0,6250
lambda (λ)	867,30	74.519,15
Ajustes al Modelo		
Pliegues de validación cruzada	5	5
Interacciones según género del Jefe de hogar	No	No
Balanceo de muestra	NA	Up-train

El modelo (12) corresponde a una regresión con la muestra original (164.960 observaciones) optimizado con una grilla de hiperparámetros donde $\alpha = \in \{0.6, 0.9\}$ y $\lambda = \in \{862.3046{,}867.3046\}$, mientras que el modelo (13) corresponde a una regresión con la muestra balanceada a través de la replicación aleatoria de las observaciones de la clase con menor representación, de manera que la muestra pasa de 164.960 observaciones a 263.872 observaciones (up-train)[9]. Si bien el modelo (13) supone una mayor precisión dentro de muestra, puede tener problemas de sobreajuste ya que el modelo falla en identificar correctamente los verdaderos negativos, por lo que no se puede calcular la especificidad. Con

[9] Es importante resaltar que la evaluación del modelo, dentro de muestra, se realizó en la base de entrenamiento original (train) y no en la base balanceada (up_train), dado que esta última toma valores creados aleatoriamente que pueden sesgar los resultados obtenidos.

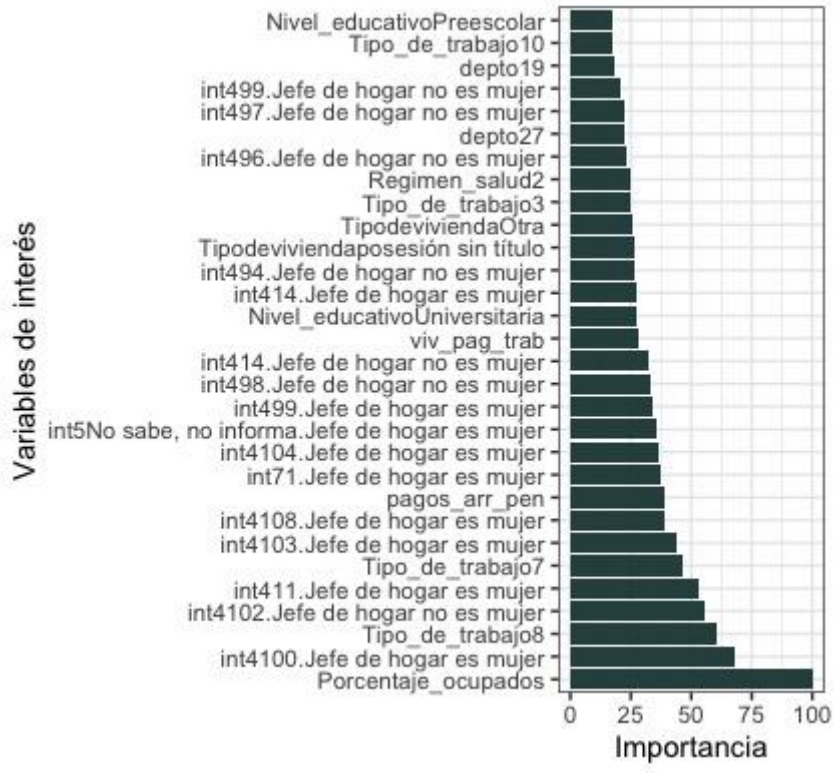
respecto a estos modelos, es importante mencionar que la distribución del ingreso predicho no presenta normalidad, como si lo hace la distribución del ingreso evidenciada en los datos (

Anexo), lo que puede suponer sesgos en la predicción.

C. Modelo final

El modelo que presentó mejor ajuste tanto dentro de muestra como fuera de muestra (test)[3], corresponde al modelo de clasificación a través de regresión logística con regularización por elastic-net (modelo 4) el cual tuvo una precisión (Accuracy) del 85,41% fuera de muestra. Este modelo toma como variables predictoras todas las características relevantes identificadas en el capítulo III de este documento; no obstante, se adelantaron los modelos (4.1) y (4.2) buscando reducir el número de predictores sin afectar la capacidad predictiva, identificando que las principales variables para la predicción son el porcentaje de ocupados, la edad, el tipo de trabajo e interacciones con las variables Jefe mujer, como se presenta en la gráfica 4:

Gráfico 4. Porcentaje de relevancia de las variables predictoras en el modelo de clasificación Logit (modelo 4).



En la Tabla 5 se presenta la comparación de los resultados de estos modelos (4), (4.1) y (4,2) junto con las especificaciones más relevantes a través de la predicción del ingreso (modelo 13) y a través de otras alternativas de clasificación (modelo 8).

Tabla 5.Comparativa de modelos con mejor ajuste

	(4)	**(4.1)**	**(4.2)**	**(8)**	**(13)**
Métricas de Desempeño					
Precisión	**0,8612**	**0,8466**	**0,8255**	0,8505	**0,9418**
Sensibilidad	0,9410	0,9344	0,9619	-	0,9418
Especificidad	0,5423	0,4957	0,2807	-	NA
Matriz de Confusión					

Verdaderos positivos (TP)	124.156	123.284	126.906		155.364
Verdaderos negativos (TF)	17.908	16.369	9.269		0
Falsos Positivos (FP)	15.116	166.55	23.755		0
Falsos Negativos (FN)	7.780	8.652	5.030		9.596
Parámetros de Regularización					
alpha (α)	0,8550	0,855	0,855	-	0,6250
lambda (λ)	0.0001	0.0005	0.0001	-	74.519,15
Ajustes al Modelo					
Corte de clasificación	0,50	0,50	0,50	0.5	NA
Pliegues de validación cruzada	10	10	10	10	5
Variables correlacionadas	No	No	No	No	No
Interacciones según genero del jefe de hogar	Sí	Sí	Sí	No	No
Balanceo de muestra	NA	NA	NA	NA	Up-train

Para el desarrollo del modelo seleccionado se utilizó una grilla de hiperparámetros donde $\alpha = \in \{0.855, 0.865\}$ y $\lambda = \in \{0.0000, 0.0005\}$ y se realizó el control de entrenamiento por validación cruzada con $k = 10$. Se identificó que los parámetros que dieron el mejor ajuste corresponden a $\alpha = 0.855$ y $\lambda = 0.0002$. Así mismo resulta relevante que aumentar el número de pliegues para validación cruzada no mejoró la capacidad predictiva.

Finalmente, respecto al desbalance de la muestra se identifica que trabajar con una muestra balanceada no mejoró la capacidad predictiva, de hecho, la empeoró al usar como criterio el clasificador de Bayes. Esto probablemente se deba a que el desbalanceo de la muestra corresponde a la distribución normal de la población, por lo que su ajuste implica que se modifique el criterio de clasificación para mejorar la predicción, tal como sucede con el modelo (5). Otra posible explicación se debe a que la base tiene un mayor número de observaciones de hogares no pobres (80%), lo cual castiga en el modelo la predicción de los hogares pobres (20%) dentro de muestra. Por esta razón, al evaluar fuera de muestra, existe el riesgo de que persista este sesgo, lo cual mejora la sensibilidad (clasificar correctamente los hogares no pobres) pero empeora la especificidad (clasificar correctamente los hogares pobres).

V. Conclusiones y recomendaciones

- Como principal conclusión no se evidencian mayores diferencias en la capacidad predictiva de los diferentes modelos desarrollados, las mejoras de ajuste se dan por la inclusión y/o exclusión de variables de interés, como es el caso de las interacciones con el género del jefe del hogar, lo que coincide con la literatura revisada. (Verme, 2020)(CEPAL, 1990)
- Las principales variables para la predicción son el porcentaje de ocupados, la edad, el tipo de trabajo e interacciones con el género del jefe del hogar.
- Si lo que se busca es poder predecir si un hogar es pobre con una alta precisión y la menor cantidad de variables, los mejores modelos son el 4.1 y 4.2, los cuales utilizan 10 y 4 variables respectivamente, sin contar las interacciones.
- Resulta relevante para la selección del modelo la menor cantidad de predictores y el menor costo computacional del modelo seleccionado.

Referencias

DANE - DNP. (2012). *Misión para el Empalme de las series de Empleo, pobreza y desigualdad.* Obtenido de https://microdatos.dane.gov.co/index.php/catalog/689/study-description

Kumar, K., Gore, A., & Sitaraman, V. (1996). Some conceptual and statistical issues on measurement of poverty. *Journal of Statistical Planning and Inference, 49*(1).

Nunes, C. (2008). Poverty Measurement: The Development of Different Approaches and its Techniques. *Society for the study of Economic Inequality*(93).

World Bank. (27 de 02 de 2018). Can State-of-the-Art Machine Learning Tools Give New Life to Household Survey Data? Obtenido de https://www.worldbank.org/en/news/feature/2018/05/30/can-state-of-the-art-machine-learning-tools-give-new-life-to-household-survey-data

World Bank. (2018). *Pover-T Tests: Predicting Poverty*. Obtenido de https://www.drivendata.org/competitions/50/worldbank-poverty-prediction/page/98/

CEPAL. (1990). The vulnerability of households headed by women: policy questions and options for latin america an the caribbean. *Meeting on vulnerable women.* Vienna: CEPAL.

Anexo 1 Procesamiento de la base de datos

A continuación, se presenta la descripción del procesamiento de datos realizado para obtener la base de datos descrita en el numeral 3 del documento:

1. Exploración de datos

Se unieron las bases ***train*** y ***test*** descargadas de la página web de la competencia en Kaggle, para después analizar cuáles variables se encontraban en la test de personas. Esto, con el fin de hacer el cruce correspondiente para aumentar el número de variables en la base test de hogares.

Después de realizar el cruce, se observa que hay una gran cantidad de variables categóricas en la base test de personas con valores omitidos (NA); específicamente, en 29 variables: *c_amortiz, oficio, tipo_oficio, aux_alim, tr_empr, prima_serv, prima_nav, prima_vac, viat_perm, bonif_anual, personas_empresa, hor_trab_seg_sem, ocup_seg_activ, dilig_trab_mas_h, disp_trab_mas_h, dilig_camb_trab, disp_camb_trab, primer_trab, ocupacion, ingr_trab, pag_pen_jub, pag_pen_alim, din_otr_per_res, din_otr_per_no_res,ayud_inst, din_prest_CDT, din_cesant, din_otr_fuent, bonif_anual*. Con el fin de realizar una mejor predicción, se decidió excluir estas variables de la base final ya que aproximadamente el 70% de los datos contenían missing values e imputarlos podría aumentar el riesgo de sesgo en las estimaciones.

2. Creación de variables

Se agregaron unas variables adicionales que pudieran explicar si un hogar es pobre o no, con los mismos datos de la muestra. En este caso, se agregaron las variables de: *porcentaje de ocupados en el hogar*, *proporción de cuántas personas del hogar hay sobre el total de cuartos del hogar* (hacinamiento) y el *promedio de los años de educación entre todos los integrantes de hogar*. Esto se hizo dado lo encontrado en la bibliografía de referencia (Ministerio de Economía y Finanzas, 2021) donde se menciona que algunas variables como el hacinamiento pueden afectar la pobreza.

3. Imputación de datos

Encontramos, para empezar, que la información correspondiente a variables de ingreso (todas ellas categóricas) tenían un porcentaje alto de missing values. Al realizar el filtro, imputamos por 0 aquellos hogares que manifestaran estar desocupados e inactivos, pues se presume que en esta condición no tienen ningún tipo de ingreso. Esto permitió reducir el porcentaje de missing en un 70%, sin embargo, al quedar un porcentaje representativo, se imputó por la moda (dato de mayor frecuencia) de cada una de estas variables*: ingreso por horas extra*, *Prima*, *Bonificación*, *Subsidio de transporte*, *Subsidio familiar*, *Subsidio de educación*, *Alimentos como pago trabajo*, *Vivienda como pago de trabajo*, *Ingresos en especie*, *Bonificación anual*, *Afiliado a fondo de pensiones*, *Otro trabajo*, *desea más horas de trabajo* y *pagos arriendo y pensiones*.

Anexo 2 Directorio de variables

Nombre	Nombre DANE	Descripción
id	Nueva	Identificación hogar
Porcentaje_ocupados	Nueva	% de ocupados en el hogar
v.cabecera	Nueva	Vive en cabecera o en la zona rural
cuartos_hog	P5000	Incluyendo sala-comedor ¿de cuántos cuartos en total dispone este hogar?
cuartos_dorm	P5010	¿en cuántos de esos cuartos duermen las personas de este hogar?
nper	Nper	Cantidad de personas por hogar
Li	Li	Línea de indigencia. Valor de la canasta básica de alimentos que establece el límite de ingresos por debajo del cual un hogar es considerado en pobreza extrema.
Lp	Lp	Valor de la canasta básica de bienes que establece el límite de ingresos por debajo del cual un hogar es considerado en pobreza.
fex_c	Fex_c	Factor de expansión
depto	Depto	codigo del departamento
fex_depto	Fex_depto	Factor Expansión departamental
d_arriendo	Nueva	dummy si el hogar paga arriendo (1) o no (0), en caso de que no es porque se trata de hogares propietarios, usufructuarios u ocupantes de hecho.
jefe_mujer	Nueva	dummy si el jefe de hogar es mujer o no
jefe_hogar	P6050	Variable que toma el valor de 1 si la observacion de hogar corresponde al jefe de hogar (en este caso, toda la info de la base hogares)
Personaxcuarto	Nueva	Proporción que indica cuantas personas del hogar hay sobre el total de cuartos del hogar (cuartos_hog/nper)
Tipodevivienda	P5090	Propia totalmente pagada, Arriendo o subarriendo, Usufructo, posesión sin título, Propia, la están pagando
Regimen_salud	P6100	1, 2, 3 pertenece al regimen contributivo o especial
Educacion_promedio	Nueva	Promedio de los años de educación entre todos los integrantes del hogar
sexo	P6020	1 hombre, 2 mujer. Sexo del jefe de hogar
edad	P6040	Años cumplidos jefe de hogar
seg_soc	P6100	¿A cuál de los siguientes regímenes de seguridad social en salud está afiliado: 1. Contributivo (eps). 2. Especial (fuerzas armadas, ecopetrol, universidades públicas) 3. Subsidiado? (eps-s) 9 . No sabe, no informa
Nivel_educativo	P6210	Media, No sabe no informa, Basica primaria, Universitaria, Ninguno, Preescolar, Basica Secundaria
Tipo_de_trabajo	P7350	1. Obrero o empleado de empresa particular. 2 Obrero o empleado del gobierno. 3. Empleado doméstico. 4. Trabajador por cuenta propia. 5. Patrón o empleador. 6. Trabajador familiar sin remuneración. 7. Trabajador sin remuneración en empresas o negocios de otros hogares. 8 Jornalero o peón. 9 Otro.

ing_hor_ext	P6510	El mes pasado recibió ingresos por concepto de horas extras? 1 sí 2 no 3 no sabe, no informa
prima	P6545	El mes pasado recibió ...: a. Primas? (técnica, de antigüedad, clima, orden público, otras, etc.) 1 si 2 no 3 no sabe, no informa
bonif	P6580	El mes pasado recibió ...: b. Algún tipo de bonificación de carácter mensual? 1 si 2 no 3 no sabe, no informa
sub_trans	P6585s2	¿cuál o cuáles de los siguientes subsidios recibió ... El mes pasado: b. Auxilio o subsidio de transporte? 1 si 2 no 3 no sabe, no informa
subsid_fam	P6585s3	¿cuál o cuáles de los siguientes subsidios recibió ... El mes pasado: c. Subsidio familiar? 1 si 2 no 3 no sabe, no informa
subsid_educ	P6585s4	¿cuál o cuáles de los siguientes subsidios recibió ... El mes pasado: d. Subsidio educativo? 1 si 2 no 3 no sabe, no informa
alim_trab	P6590	Además del salario en dinero, ¿el mes pasado recibió alimentos como parte de pago por su trabajo? 1 si 2 no 3 no sabe, no informa
viv_pag_trab	P6600	Además del salario en dinero, ¿el mes pasado recibió vivienda como parte de pago por su trabajo? 1 si 2 no 3 no sabe, no informa
ing_esp	P6620	Además del salario en dinero, ¿el mes pasado... Recibió otros ingresos en especie por su trabajo(electrodomésticos, ropa, productos diferentes a alimentos o bonos tipo sodexho)? 1 si 0 no 3 no sabe, no informa
bonif_anual	P6630s6	En los últimos 12 meses recibió: e. bonificaciones anuales 1 sí 2 no
fondo_pensiones	P6920	Está... Cotizando actualmente a un fondo de pensiones? 1 sí 0 no
otro_trab	P7040	Además de la ocupación principal, ¿.... tenía la semana pasada otro trabajo o negocio? 1 sí 0 no
hor_trab_sem	P6800	¿cuántas horas a la semana trabaja normalmente.... en ese trabajo ?
deseo_hor	P7090	Además de las horas que trabaja actualmente ¿..... quiere trabajar más horas? 1 sí 0 no
pagos_arr_pen	P7495	El mes pasado, ¿recibió pagos por concepto de arriendos y/o pensiones? 1 sí 0 no
din_otr_per	P7505	Durante los últimos doce meses, ¿recibió dinero de otros hogares, personas o instituciones no gubernamentales; dinero por intereses, dividendos, utilidades o cesantias? 1 sí 0 no
pet	Pet	Población en edad de trabajar 1: sí 0: no
ocupado	Oc	Ocupado 1: sí. 0: no
desocupado	Des	Desocupado 1: sí. 0: no
inactivo	Ina	Inactivo 1: sí. 0: no
Pobre	Pobre	Variable que idenfifica los hogares en condiciones de pobreza. Pobre= 1 sí, 0= No
IngresoPerCapita	ingpcug	Ingreso percápita de la unidad de gasto con imputación de arriendo a propietarios y usufructuarios

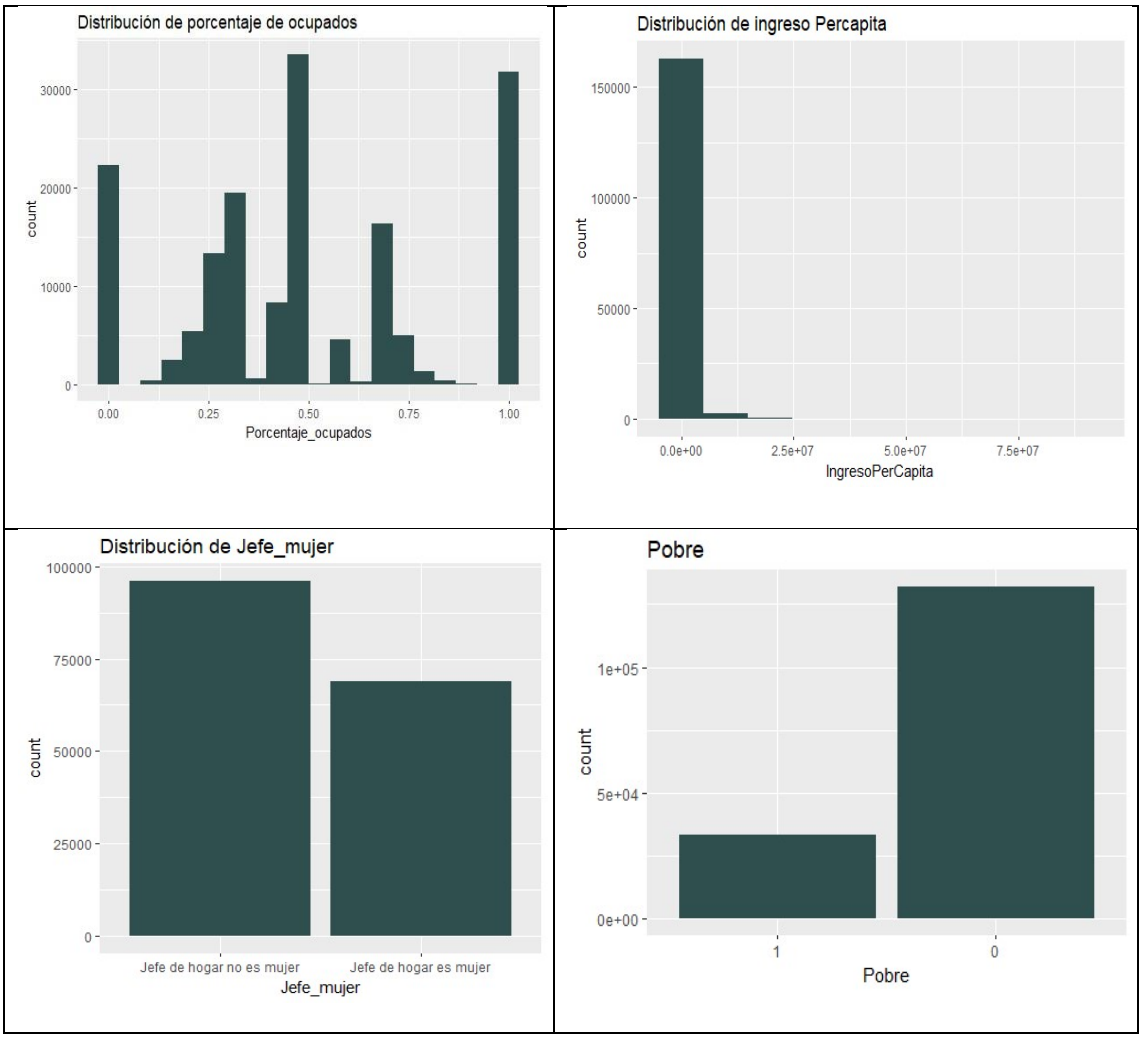
Anexo 3 Otros análisis y gráficos

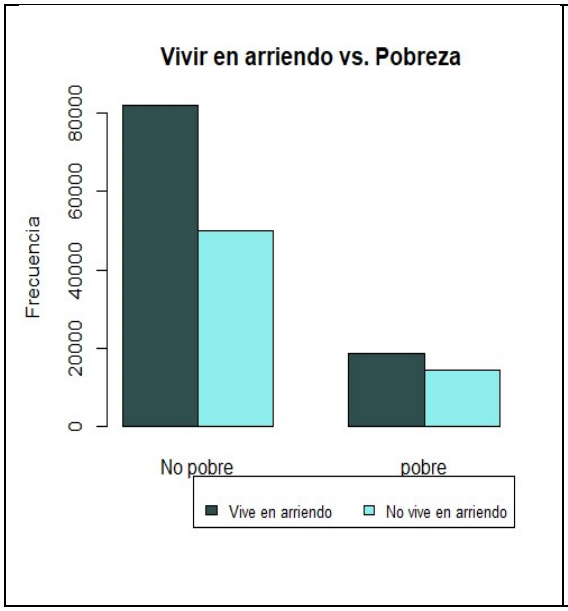
Tabla 3.1. Estadísticas descriptivas de variables categóricas

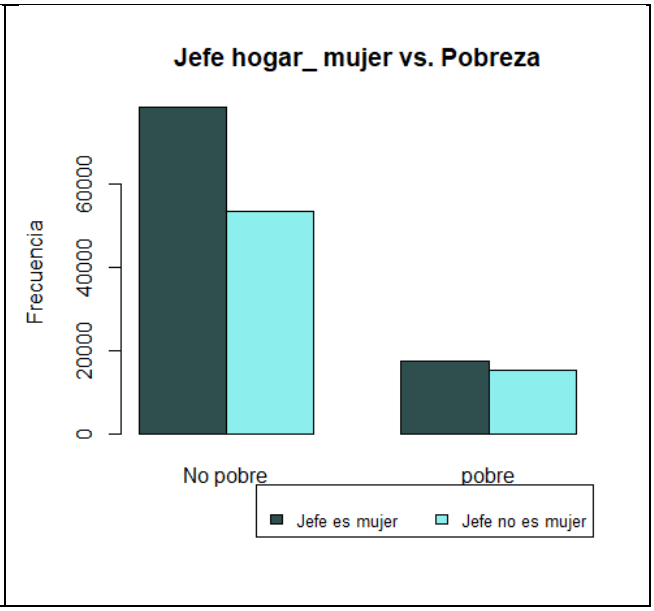
	Test				Train			
	N. obs.	**Sí**	**No**	**No sabe/no responde**	**N. obs.**	**Sí**	**No**	**No sabe/no responde**
El hogar es en arriendo	66.168	61,9%	38,1%	0,00%	164960	61%	39,1%	0,00%
Jefe hogar mujer	66.168	41,6%	58,4%	0,00%	164960	42%	58,2%	0,00%
Ingreso por horas extra	66.168	2,2%	97,8%	0,03%	164960	2%	97,6%	0,03%
Prima	66.168	0,6%	99,4%	0,03%	164960	1%	99,4%	0,03%
Bonificacion	66.168	1,1%	98,9%	0,03%	164960	1%	98,8%	0,03%
Subsidio de transporte	66.168	13,8%	86,1%	0,09%	164960	14%	85,6%	0,09%
Subsidio familiar	66.168	7,8%	92,1%	0,03%	164960	8%	91,9%	0,04%
Subsidio de educacion	66.168	0,1%	99,8%	0,01%	164960	0%	99,8%	0,01%
Alimentos como pago trabajo	66.168	3,9%	96,1%	0,01%	164960	4%	96,0%	0,01%
Vivienda como pago de trabajo	66.168	1,4%	98,6%	0,01%	164960	1%	98,7%	0,01%
Ingresos en especie	66.168	0,3%	99,7%	0,03%	164960	0%	99,7%	0,02%
Bonificación anual	66.168	0,1%	99,9%	0,00%	164960	0%	99,9%	0,00%
Afiliado a fondo de pensiones	66.168	0,1%	99,9%	0,00%	164960	0%	99,9%	0,00%
Otro trabajo	66.168	4,2%	95,8%	0,00%	164960	4%	95,9%	0,00%
Desea mas horas de trabajo	66.168	5,6%	94,4%	0,00%	164960	6%	94,2%	0,00%
Pagos arriendo y pension	66.168	17,0%	83,0%	0,00%	164960	18%	82,3%	0,00%
Dinero de otras personas	66.168	28,0%	72,0%	0,00%	164960	29%	71,4%	0,00%

Jefe de hogar ocupado	66.168	71,1%	28,9%	0,00%	164960	71%	29,0%	0,00%
Jefe de hogar desocupado	66.168	4,6%	95,4%	0,00%	164960	5%	95,3%	0,00%
Jefe de hogar inactivo	66.168	24,4%	75,6%	0,00%	164960	24%	75,7%	0,00%
Pobre	0	0%	0%	0,00%	164960	20%	80,0%	0,00%

Gráfico A3.2. Otros gráficos del análisis descriptivo




Vivir en arriendo vs. Pobreza
Frecuencia
0
20000
40000
60000
80000
No pobre
pobre
Vive en arriendo
No vive en arriendo


Jefe hogar_ mujer vs. Pobreza
Frecuencia
0
20000
40000
60000
No pobre
pobre
Jefe es mujer
Jefe no es mujer

Anexo 5.

Distribucion del ingreso real frente al estimado.