IDENTIFICACIÓN DE PATRONES DE CONSUMO DE LOS VENEZOLANOS MEDIANTE MÁQUINAS DE VECTORES SOPORTE

Liz J. Aranguren P.
Gerardo Colmenares

RESUMEN

El estudio realizado estuvo dirigido a identificar patrones de consumo de los venezolanos utilizando Máquinas de Vectores Soporte (MVS) a partir de datos obtenidos de la II Encuesta Nacional de Presupuestos Familiares (1997-1998). La importancia de la investigación radica en la utilización de la técnica de análisis no lineal antes mencionada, la cual se caracteriza por la incorporación del principio de Minimización de Riesgo Estructural, que en algunos casos ha demostrado ser superior al principio tradicional de Minimización de Riesgo Empírico, empleados por redes neuronales y otros métodos lineales convencionales. La técnica multivariante Análisis de Componentes Principales, utilizada con fines exploratorios, permitió definir las cuatro variables latentes referentes a las modalidades de consumo de los venezolanos. Dichas variables fueron usadas para establecer la variable de salida para la clasificación mediante las MVS. El estudio permitió afirmar que la metodología empleada con Máquinas de Vectores Soporte se desempeña muy bien en estudios sobre identificación de patrones de consumo, arrojando errores mínimos en las clasificaciones. Entre las conclusiones se destaca que dicha técnica produce un modelo consistente y mostró gran capacidad para generalizar, incluso cuando el entrenamiento se hizo a través de pocos ejemplos. Además, se concluye la importancia de los modelos híbridos obtenidos mediante la unión de diferentes áreas del conocimiento, como los son la estadística aplicada y el aprendizaje automático.

Palabras clave: Patrones de Consumo, Encuesta de Presupuestos Familiares, Análisis de Componentes Principales, Máquinas de Vectores Soporte

1. INTRODUCCIÓN

En la era de la globalización y de los continuos avances científicos y tecnológicos, de los cuales no escapa Venezuela, las prioridades y las metas de la sociedad se encuentran en constante cambio, lo que hace esencial detectar y pronosticar esos cambios a través de estudios relacionados con diferentes estándares de vida, en general, y sobre consumo, en particular. Los estudios sobre patrones de consumo

permiten explicar las características del desarrollo humano, el nivel de vida, los cambios en los estilos de vida, entre otros aspectos. En un sentido bastante amplio, los patrones de consumo familiares son indicadores a tomar en cuenta en miras a explicar el proceso de desarrollo humano, el cual puede ser entendido como el proceso de ampliación de oportunidades para las personas.

Los datos resultantes de la II Encuesta Nacional de Presupuestos Familiares (ENPF) aplicada en los años 1997- 1998, fueron utilizados en la presente investigación para identificar los patrones de consumo de los venezolanos mediante Máquinas de Vectores Soporte (MVS) como técnica de clasificación no lineal. Gunn (1998) afirma que las MVS están ganando popularidad debido a sus tantas características atractivas, entre las cuales se destaca la incorporación del principio de Minimización de Riesgo Estructural, el cual ha demostrado ser superior al principio tradicional de Minimización de Riesgo Empírico, empleados por Redes Neuronales y otros métodos lineales convencionales. Este autor explica que el principio de Minimización de Riesgo Estructural minimiza un límite superior de riesgo esperado, opuesto al principio de Minimización de Riesgo Empírico que minimiza el error en los datos de entrenamiento. Esta diferencia dota a las MVS con una mayor habilidad para generalizar, es decir, tener una alta capacidad de pronóstico para nuevas observaciones.

Investigaciones previas sobre presupuestos familiares, realizadas por Elsy Garnica en el año 1996; y Anido, Orlandoni, y Quintero, en el año 2005, entre otras, tomando como base los datos provenientes de Encuestas Nacionales de Presupuestos Familiares, se condujeron con la utilización de técnicas de análisis lineales, destacándose que las mismas estuvieron delimitadas a la ciudad de Mérida. En la investigación que se presenta, además de incursionar con las Máquinas de Vectores Soporte como técnica de clasificación no lineal, se utilizaron datos a nivel nacional, provenientes de las II Encuestas Nacionales de Presupuestos Familiares, tal como se ha señalado anteriormente.

2. PREPROCESAMIENTO DE LOS DATOS

Comúnmente, los datos originales con los que se desea trabajar contienen datos u observaciones incompletos (valores de algunos atributos faltantes) y/o datos inconsistentes. Precisamente, uno de los pasos dentro del preprocesamiento es encontrar los valores atípicos, conocidos comúnmente como *outliers*, y, por otro lado, descubrir valores faltantes dentro de la muestra de datos. La importancia de esto se debe a que la presencia tanto de valores faltantes como de valores atípicos puede llevar a la extracción de patrones poco útiles o errados; mientras que, un conjunto de datos formado solamente por valores consistentes y completos permite una extracción de patrones de

calidad, acertados y útiles. Para el caso particular del estudio realizado, el preprocesamiento permitió, entre otras cosas, reducir el tamaño de la muestra y mejorar la calidad de los datos contenidos en ella para la aplicación de la técnica multivariante Análisis de Componentes Principales.

2.1 ESTRUCTURA ORIGINAL DE LA MUESTRA

Originalmente, los datos obtenidos para la realización del estudio se encontraban en una base de datos en el programa Microsoft Access, la cual fue diseñada por Márquez (2002) como parte del trabajo de grado para optar al título de Magíster en Estadística. Dicha base de datos se encuentra formada por 23 tablas; en las cuales estaba distribuida información cualitativa y cuantitativa tanto de los hogares que fueron encuestados como de las viviendas en la que ellos habitaban para ese momento con distinta información respecto a los resultados de la II Encuesta Nacional de Presupuestos Familiares.

El estudio minucioso de cada una de las tablas permitió concluir que las que contenían información relevante para el estudio eran: entidades, gastos diarios del hogar, gastos diarios personales, gastos del hogar en restaurantes, gastos mensuales, gastos personales en restaurantes, gastos trimestrales y anuales, ingresos, región y vivienda03. Este hecho se justifica con las variables que se eligieron para llevar a cabo la identificación de patrones de consumo familiares, las cuales fueron: los desembolsos de los hogares en cada uno de los grupos de gastos que se establecerían y el ingreso total familiar.

2.2 DETERMINACIÓN DE LOS GRUPOS DE GASTOS

Luego de haberse familiarizado bien con los datos originales y de haber determinado dónde y cómo se encontraba específicamente la información requerida para el estudio, fue necesario determinar los grupos de gastos que iban a ser considerados. En este punto, se profundizó en la agrupación que los gastos tenían inicialmente, los cuales se hallaban clasificados en nueve grupos, a saber: 10) Alimentos, bebidas y tabaco; 20) Vestido y calzado; 30) Gastos de vivienda y sus servicios; 40) Mobiliario, equipos del hogar y mantenimiento de la vivienda; 50) Salud; 60)Transporte y comunicaciones; 70) Educación, cultura y esparcimiento; 80)Artículos, efectos personales y servicios diversos; y, por último, 90) Otros gastos.

Aunque la clasificación inicial de los datos pueda resultar suficientemente lógica, se decidió que ésta debía ser modificada. La razón que fundamenta una nueva clasificación tiene que ver con el objetivo principal del estudio que era identificar patrones de consumo de los venezolanos. La comprensión de la composición original de cada uno de los grupos permitió establecer que era necesario desglosar algunos de ellos ya que resultaban bastante extensos

respecto al tipo de productos y servicios que estaban incluidos en ellos. Los 16 grupos de gastos definitivos quedaron establecidos de la siguiente manera: 1) Alimentos y bebidas no alcohólicas, 2) Bebidas alcohólicas y tabaco, 3) Alimentos y bebidas tomadas fuera del hogar, 4) Vestido y calzado, 5) Vivienda y sus servicios, 6) Mobiliario y equipos del hogar, 7) Mantenimiento de la vivienda, 8) Salud, 9) Transporte, 10) Comunicación, 11) Educación y cultura, 12) Recreación y esparcimiento, 13) Artículos, efectos personales y servicios diversos, 14) Gastos financieros, tributarios y legales, 15) Viajes y 16) Otros gastos.

2.3 CREACIÓN DE LA MATRIZ DE DATOS

Una vez establecidos los grupos de gastos con los que se trabajaría se pudo proceder a realizar la matriz de datos, la cual permitió un manejo fácil para las posteriores aplicaciones de las técnicas ACP y MVS. Se determinó que dicha matriz debía estar compuesta por 20 columnas en total y estaría estructurada como se muestra en la tabla 1. Sin embargo, para llegar a tenerla de esa manera fue necesario realizar varios procedimientos, los cuales se describirán a continuación. Para crear la matriz de datos se totalizaron los gastos por grupos y los ingresos para cada una de las familias estudiadas y, seguidamente, se unificó toda la información para obtener los datos organizados como se muestra en la siguiente tabla. Luego de esto fue necesario hacer el manejo de los valores faltantes y de los valores atípicos.

NUMERBCV	COD_ENTIDAD	NUM_MIEMBROS	GASTO G1	GASTO G2	…	GASTO G16	INGRESO
XXXXXX	XX	XX	XXXXX	XXXXX	…	XXXXX	XXXXX
XXXXXX	XX	XX	XXXXX	XXXXX	…	XXXXX	XXXXX
XXXXXX	XX	XX	XXXXX	XXXXX	…	XXXXX	XXXXX
XXXXXX	XX	XX	XXXXX	XXXXX	…	XXXXX	XXXXX

Tabla 1: Estructuración de matriz de datos

Cuando se creó la matriz de datos se pudo observar claramente que existía una gran cantidad de valores faltantes los cuales debían ser atendidos y manejados de diferentes maneras. Los “huecos” relacionados con los gastos se rellenaron con ceros ya que esta falta de valor indicaba que el hogar no registró gastos en bienes o servicios pertenecientes al grupo donde no estaba indicado ningún valor. Sin embargo, esta forma de manejar datos ausentes no se pudo aplicar a las otras columnas debido a que no resultaba lógico ajustar su valor a cero. Por ejemplo, no se podía considerar que un hogar sin ingreso especificado tenía un ingreso igual a cero, pues en un mes es necesario que un hogar obtenga de cualquier manera un ingreso aunque sea mínimo para poder

incurrir en gastos. De esta manera, el manejo de datos ausentes respecto a los ingresos fue la eliminación de aquellos registros que no especificaron ningún ingreso. Para la información en cuanto a la entidad y al número de miembros de cada hogar se hizo de manera análoga a la de los ingresos. Esto se justifica en el hecho de que se estudió el consumo sólo para un momento determinado y no para una serie de tiempos, por lo que no existía manera de obtener esa información a través de resultados de la misma encuesta aplicada en años distintos. Después de haber manejado los datos ausentes se logró obtener una matriz de 8406 observaciones, todas con datos completos.

Seguidamente del manejo de datos faltantes se pasó a estudiar detenidamente cada uno de las variables incluidas. Inicialmente, se realizaron análisis a los gastos con la intensión de observar sus valores mínimos y máximos, dispersiones, varianzas y medias. Esta parte se trató con *Statgraphics*, un programa estadístico que permite realizar una gran variedad de análisis estadísticos a grandes volúmenes de datos. Particularmente, el procedimiento que permitió realizar los análisis mencionados fue el "Análisis Unidimensional", especificado dentro del manejo de datos numéricos en la barra del menú "Descripción". Este procedimiento, al permitir observar la dispersión de los valores, hace que sea fácil detectar gráficamente aquellos valores que sean atípicos. El análisis unidimensional se le realizó a todos los gastos y a los ingresos considerando que son las variables de estudio, ya que aquella información respecto a características de los hogares (entidad de ubicación y número de miembros) no son variables de estudio sino que aportan información adicional que permite realizar distintos análisis posteriores.

El análisis de dispersión de cada uno de los grupos de gastos por separado permitió determinar que existían cuatros valores atípicos indiscutibles, de los cuales dos se referían a montos en el grupo 8 (Salud), uno en el grupo 18 (Gastos financieros) y otro en el grupo 19 (Gastos tributarios y legales). Sin importar cual haya sido la razón de esos valores, se decidió eliminarlos de los datos debido a que podían influir seriamente en los resultados. Para incurrir en la eliminación de los cuatro registros asociados a esos cuatro valores se decidió sustituirlos por las medias de cada grupo al que pertenecían. Por ejemplo, los dos valores altos del grupo de salud se sustituyeron por la media de ese mismo grupo.

De esta manera, se concluyó la fase del preprocesamiento, la cual sin ninguna duda, forma parte fundamental de todo estudio relacionado con minería de datos. Este paso resultó muy útil por cuanto permitió conocer bien los datos de la II ENPF, darles forma y coherencia, permitiendo que los mismos pudiesen ser manipulados para lograr el objetivo buscado, el cual era conseguir un modelo óptimo de MVS para identificar patrones de consumo de las familias venezolanas.

3. EXPERIMENTACIÓN Y ANÁLISIS DE RESULTADOS

3.1 ANÁLISIS DE COMPONENTES PRINCIPALES

Una vez obtenida la matriz de datos sin valores atípicos gráficamente evidentes y sin valores faltantes, se pudo proceder a realizar el análisis de componentes principales, para lo cual se recurrió al programa estadístico *Statgraphics 5.1*. Dicho programa desarrollado por la compañía norteamericana Manugistics, Inc., "fue diseñado para el análisis estadístico de datos con el objetivo de resolver problemas de estadística descriptiva, inferencial o ambos" (Lozano, 2008, p. 1). Entre las características atractivas de *Statgraphics* se destaca su fácil manipulación y sus grandes capacidades gráficas.

Cabe destacar que resultó necesario realizar diversas corridas para determinar cuál era la que arrojaba mejores resultados en cuanto a la variabilidad percibida de los datos originales y a las variables representadas por cada uno de los componentes principales. Además, es fundamental resaltar que el ACP fue utilizado con fines exploratorios y fue el que permitió reducir la dimensionalidad de los datos originales. Asimismo, esta técnica permitió identificar subconjuntos de variables relacionadas directamente con la manera cómo consumían los venezolanos para el momento en el que se aplicó la encuesta.

3.1.1 Cálculo y selección de componentes principales

Primero, se realizaron varias corridas utilizando la matriz de varianza-covarianza con la intención de verificar la tendencia de que aquellas variables con grandes varianzas, en comparación con las de las demás variables, figurarían dentro de los primeros componentes. Las corridas consistieron en aplicar la técnica para los datos originales y para los datos escalados (entre 0 y 1). Los resultados obtenidos fueron los esperados debido a que los gastos financieros, tributarios y legales figuraban dentro de los componentes con mayores pesos de importancia, es decir los primeros componentes. Además, se observó en los gráficos que los componentes resultaban bastante difíciles de interpretar debido a que la mayoría de las variables se encontraban aglomeradas, mientras que aquellas con sus varianzas más elevadas eran las que se distinguían de las demás.

De esta manera, se procedió a aplicar la técnica utilizando la matriz de correlación, debido a que se observó grandes diferencias en las varianzas de las distintas variables. El uso de la matriz de correlación para el ACP permite darle la misma

importancia a todas las variables y, con esto, se evita que haya alguna inclinación hacia aquellas variables con varianzas considerablemente más altas. Es importante tener presente que la intención de la aplicación del ACP fue: primero, reducir el número de variables a un menor número de variables latentes que permitirían determinar las clases que se utilizarían para la clasificación con MVS; y segundo, la selección de componentes principales permitiría determinar grupos de variables asociadas o correlacionadas entre sí, con los cuales se puede establecer maneras de consumo de los venezolanos. De igual manera que se hizo con la matriz de varianza-covarianza, se hicieron corridas tanto para los datos escalados como para los datos originales. No obstante, se observó que ambos arrojaban resultados muy similares en cuanto a la cantidad de componentes principales y la composición de cada uno.

En la tabla 2 puede observarse los resultados obtenidos de la última corrida, realizada a través de la matriz de correlación para los 16 grupos de gastos establecidos. En cuanto a la variabilidad de los datos originales, estos resultados no se consideran muy buenos ya que cuatro componentes explican el 50,45% de la misma. Sin embargo, los resultados obtenidos en cuanto a la composición de los componentes se consideraron satisfactorios y, como punto muy importante, se observó que la mayoría de las variables originales se destacaban dentro de esos cuatro componentes. Tomando en cuenta que el objetivo era reducir las variables originales a un nuevo número menor de variables latentes que permitieran definir las clases necesarias, se seleccionaron los resultados de la última corrida para analizar los componentes principales y, posteriormente, identificar patrones de consumo de los venezolanos a través de las MVS. En el gráfico 1 se puede observar los pesos de los autovalores de todos los componentes y la recta que indica el número de componentes principales a seleccionar.

Componente Número	Porcentaje de varianza	Porcentaje acumulado
1	24,10%	24,10%
2	7,50%	31,60%
3	6,39%	37,99%
4	**6,23%**	**50,45%**

Tabla 2: Resultados de la corrida final del ACP

3.1.2 Análisis de la matriz factorial

Luego de haber realizado el procedimiento necesario para el cálculo de los componentes y de haber seleccionado el número de componentes principales, el siguiente paso realizado fue el análisis de la matriz factorial. Para esto, es necesario estudiar tanto la magnitud como el signo de las correlaciones obtenidas a través de la aplicación de la técnica. Tomado esto en cuenta, se procedió a analizar cada uno de los

cuatro componentes principales obtenidos a través de la matriz de correlación, utilizando como variables iniciales los 16 grupos de gastos definitivos. Los pesos de los componentes se muestran en la tabla 3.

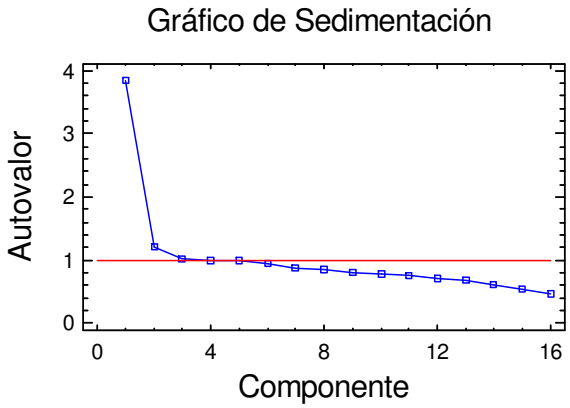

Figura 1: Resultados gráficos de la corrida final del ACP

Para el caso del primer componente, se pudo observar que el mismo tiene una correlación positiva alta con Gasto Grupo 5, Gasto Grupo 10, Gasto Grupo 9 y Gasto Grupo 13. Además tiene correlaciones positivas menores pero aún considerables (cerca de 0,3) con Gasto Grupo 3, Gasto Grupo 11, Gasto Grupo 1 y con Gasto Grupo 4. Este factor, el cual explica un gran porcentaje de la variabilidad de los datos, se puede interpretar como aquellos venezolanos que tienden a tener gastos elevados en vivienda y sus servicios, comunicación, transporte, efectos personales, comidas fuera del hogar, educación y cultura, alimentos y bebidas no alcohólicas, vestido y calzado.

	Componente 1	**Componente 2**	**Componente 3**	**Componente 4**
Gasto Grupo 1	**0,260604**	**-0,441452**	0,0870371	-0,183408
Gasto Grupo 2	0.169713	**-0.518719**	**0.353658**	0.197392
Gasto Grupo 3	**0.280332**	0.0182913	0.186094	0.180815
Gasto Grupo 4	**0.251819**	**-0.286009**	-0.191973	0.0526567
Gasto Grupo 5	**0.353329**	**0.284436**	0.0193821	0.000429061
Gasto Grupo 6	0.183823	0.012598	**-0.460809**	-0.164882
Gasto Grupo 7	0.162907	-0.13792	-0.216474	**-0.61749**
Gasto Grupo 8	0.1645	0.179716	0.0412534	-0.200087
Gasto Grupo 9	**0.339314**	0.078296	0.0142412	-0.0440675
Gasto Grupo 10	**0.345508**	0.238121	0.0349837	0.0391123
Gasto Grupo 11	**0.280097**	**0.257896**	-0.0685882	-0.0258961
Gasto Grupo 12	0.236216	-0.0598217	-0.124448	**0.255883**
Gasto Grupo 13	**0.311634**	**-0.292153**	-0.0277551	-0.0215409
Gasto Grupo 14	0.0236996	-0.084531	**-0.642927**	**0.530986**
Gasto Grupo 15	0.192405	0.141972	**0.30435**	**0.302191**
Gasto Grupo 16	0.21353	**0.275601**	0.0788011	-0.00820293

Tabla 3: Matriz de pesos de los componentes

Por otra parte, el componente 2 tiene mayor correlación positiva con Gasto Grupo 5, Gasto Grupo 16 y Gasto Grupo 11, mientras que tiene correlación negativa

con Gasto Grupo 2, 1, 13 y 4. Este componente se refiere a aquellas familias que le dan prioridad a gastos en vivienda y sus servicios, educación, cultura y otros gastos. Además, se puede decir que la correlación negativa mencionada, muestra que los individuos que se comportan de esta manera tienden a tener bajos desembolsos en bebidas alcohólicas, tabaco, alimentos y bebidas no alcohólicas, efectos personales, vestido y calzado.

El tercer componente muestra una alta correlación positiva con Gasto Grupo 2 y con Gasto Grupo 15 y, a su vez, correlación negativa alta con gastos en los grupos 14 y 6. Los venezolanos que actúan de acuerdo a este componente le dan gran prioridad a gastos en viajes, bebidas alcohólicas y tabaco. Sin embargo, muestran bajos desembolsos o ninguno en gastos financieros, tributarios y legales y en mobiliario y equipos del hogar. Los gastos en viajes contemplan pagos realizados en pasajes terrestres, marítimos y aéreos al exterior o al interior del país, alquiler de vehículos de transporte, hospedaje en hoteles y posadas, etc.

El componente 4 tiene una correlación positiva alta con los grupos de gastos identificados como 14, 15 y 12. Asimismo, presenta una correlación negativa alta con Gasto Grupo 7. Esto se pudo interpretar como el alto consumo en gastos financieros, tributarios y legales, viajes, diversión y esparcimiento. Los venezolanos que presentan este tipo de comportamiento respecto a su consumo, también muestran muy bajos desembolsos, incluso nulos, en gastos relacionados con el mantenimiento de la vivienda. Algunos de los gastos que allí se incluyen son desembolsos en servicios bancarios (compra de cheques de gerencia, emisión de tarjetas de crédito, etc.), notarías, consumos con tarjetas de crédito, pagos de impuestos e intereses, entre otros.

3.1.3 Interpretación de los componentes principales

Uno de los pasos claves del Análisis de Componentes Principales es la interpretación de los factores o componentes principales, la cual debe realizarse a través de la observación de relaciones de dichos factores con las variables originales y entre ellos mismos. Este paso jugó un papel fundamental en el desarrollo del proyecto, debido a que la definición de las clases que se establecerían para realizar la clasificación con las MVS, dependía estrechamente de la interpretación que se le diera a los componentes principales seleccionados. En otras palabras, las clases requeridas se definirían a partir de las variables latentes extraídas del ACP.

Conocimientos de los expertos en la materia es sumamente necesario en el momento de la interpretación de los componentes principales de un determinado estudio. Por lo tanto, para llevar a cabo esta parte del estudio se recurrió a un experto en la materia, el cual determinó necesario eliminar de los componentes las variables que

resultaban ambiguas. Esto se hizo con la finalidad de que las variables en los componentes fueran completamente excluyentes, lo cual permitiría definir con mayor facilidad las nuevas variables. Para ello, dicho experto hizo un estudio de las correlaciones de las variables y de la definición de cada uno de los componentes. Luego de indagar en la combinación de las variables obtenidas, el mismo sugirió nombrar las cuatro nuevas variables como se señala en la tabla 4, en la cual se puede observar, de manera resumida, las variables que contemplan cada uno de los componentes principales y el significado de cada uno de ellos. Con esto, se puede percibir de manera más clara la definición de los cuatro ejes correspondientes a los cuatro componentes seleccionados, los cuales permitirían definir las cuatro maneras prevalecientes de consumo de los venezolanos.

CP	Variables	Significado	Interpretación
1	g5, g10, g9, g13, g3, g1, g4	Vivienda y sus servicios, comunicación, transporte, efectos personales, comidas fuera del hogar, alimentos y bebidas no alcohólicas, vestido y calzado	Gastos básicos
2	g11, g16	Educación, cultura y otros gastos	Gastos de educación y servicios diversos
3	g2, g15, g12	Bebidas alcohólicas, tabaco, viajes, diversión y esparcimiento	Gastos de recreación
4	g14	Gastos financieros, tributarios y legales	Gastos financieros, tributarios y legales

Tabla 4: Interpretación y composición de las variables definidas

De acuerdo a los resultados obtenidos del Análisis de Componentes Principales y a la interpretación que se le dio a los mismos, se puede decir que la mayoría de los venezolanos dan prioridad a gastos básicos, los cuales se refieren a vivienda y sus servicios, alimentación, vestido y calzado, transporte (público y privado), comunicación (telefonía fija, móvil, Internet, etc.) y efectos personales. Por lo tanto, los mismos gastan la mayoría o todo su dinero en los rubros mencionados y el restante en los demás rubros. Vivienda y sus servicios incluyen gastos en alquileres de inmuebles, servicios domésticos, seguros relacionados con la vivienda y servicios como agua, electricidad, gas, etc. Los efectos personales abarcan todos los productos y servicios de higiene, cuidado y apariencia personal como cepillos de diferentes tipos, cortes de cabello, jabones, máquinas de afeitar, secadores de pelo, maquillaje, prendas, relojes, etc.

Por otro lado, existe una menor porción de la población que destina la mayoría de sus ingresos a gastos de educación y servicios diversos. Los gastos en educación y/o cultura contemplan costos en artículos escolares, computadoras, enciclopedias, libros,

instrumentos musicales, clases particulares, matrícula en educación básica, superior, extracurricular, etc. Los servicios diversos incluyen desembolsos en productos y servicios como matas, animales, donaciones, copias fotostáticas y de llaves, servicios funerarios, armas de fuego, entre otros.

Asimismo, hay quienes le dan prioridad a gastos de recreación, los cuales se refieren a pagos en entradas a lugares de esparcimiento (parques de diversiones, museos, etc.), cámaras fotográficas, equipos de música, equipos y artículos deportivos, lotería, apuestas de caballos y gastos relacionados con viajes (hospedaje en posadas y hoteles, pasajes terrestres, aéreos y marítimos al interior y exterior del país, tours, etc.). Además, los gastos de recreación incluyen gastos en todo tipo de bebidas alcohólicas y tabaco.

En contraparte a las tendencias de consumo antes mencionadas, existe una pequeña porción de venezolanos que tienen altos desembolsos en gastos financieros, tributarios y legales. Algunos de los gastos incluidos en este grupo son: servicios financieros (emisión de chequeras, tarjetas de crédito y otros servicios bancarios), gastos en notarías y registros, honorarios profesionales por servicios jurídicos, pagos de impuestos e intereses sobre préstamos, etc.

3.2 MÁQUINAS DE VECTORES SOPORTE

La aplicación de la técnica Análisis de Componentes Principales, la cuál se utilizó con fines exploratorios, permitió definir cuatro clases que se utilizaron, posteriormente, para realizar multiclasificación con Máquinas de Vectores Soporte. Dichas clases a saber son: Gastos básicos (clase 1), Gastos de educación y servicios diversos (clase 2), Gastos de recreación (clase 3) y Gastos financieros, tributarios y legales (clase 4). De manera general, la clasificación con MVS consistió en determinar para las familias estudiadas su tipo de consumo; es decir, a cuál de las cuatro clases pertenecía, dependiendo de la distribución de sus gastos. Para esto, se recurrió al *software Weka 3.4.12*, el cual fue desarrollado por la Universidad de Waikato en Nueva Zelanda.

3.2.1 Selección de las muestras

El primer paso realizado para lograr la clasificación de los venezolanos de acuerdo a su modalidad de consumo fue la selección de las muestras, las cuales se seleccionaron a través de muestreo estratificado aleatorio. En este caso particular, los estratos que se consideraron fueron las 23 entidades del país debido a que se quiso mantener la idea de inclusión de todas las regiones del país estudiadas por las encuestas. De esta manera, lo que se hizo fue determinar la proporción de observaciones que tenía

la muestra total de los datos para cada entidad y, luego, se extrajeron muestras aleatorias simples de cada entidad manteniendo las mismas proporciones. Para esto, primero se determinó el tamaño de la muestra a partir de la fórmula comúnmente conocida cuando se conoce el tamaño de la población.

Conociendo la fórmula que se debía utilizar y los parámetros necesarios para su estimación, se pudo proceder a realizar el cálculo del tamaño de la muestra con la que se realizaría el entrenamiento de las Máquinas de Vectores Soporte. Para esto, se tomó en cuenta que se quería realizar una estimación con seguridad de 95% y, con respecto a la precisión, se decidió probar el desempeño de las MVS para un error máximo de estimación de 3% y de 1%. Los diferentes valores de precisión permitirían observar el comportamiento de las MVS respecto a diferentes tamaños de muestra para el entrenamiento y, con esto, se evaluaría su habilidad para generalizar cuando se entrenaba con muestras relativamente grandes y con muestras mucho más pequeñas. En relación al valor de *p*, luego de revisar estudios relacionados y determinar que no se deseaba crear muestras muy grandes debido al pequeño tamaño de la muestra original, se decidió utilizar un valor de 0,05. De esta manera, los dos tamaños de muestras determinados son de 1500 (error permitido de 0,03) y 198 (error permitido de 0,01).

Luego de calculados los tamaños de muestra que se utilizarían, fue necesario calcular la proporción de observaciones que cada entidad aportaba a la muestra total y, en función de esas proporciones, se extrajeron, a través de muestreo aleatorio simple, las cantidades necesarias de cada entidad para conformar la muestra de entrenamiento.

Conocido estos valores y con la idea de realizar varias pruebas a las distintas configuraciones de MVS que se propondrían y luego estudiar la generalización de aquella configuración que se considerará óptima, se procedió a extraer diferentes muestras aleatorias para cada entidad. La extracción de muestras aleatorias sin reposición se realizó mediante *Matlab 7.0*, a través de un programa realizado con ese fin, el cual permitía generar cierta cantidad de números aleatorios sin repetición, para un cierto rango de números ingresados por el usuario. *Matlab* es un programa de cálculo técnico y científico, el cual tiene un lenguaje propio y, además, dispone de un código básico y varias librerías especializadas. Previo a esto, fue necesario ordenar las observaciones de la muestra total de acuerdo a las entidades, lo que permitió definir un rango específico para cada una de ellas.

Por otra parte, el programa *Weka,* el cual se decidió utilizar para la clasificación, brinda diferentes opciones para la selección de muestras para el entrenamiento y la prueba, las cuales pueden utilizarse con la muestra total de los datos. La primera, se refiere a la utilización de toda la muestra para el entrenamiento y la misma para la prueba. La segunda de las opciones que ofrece *Weka* se refiere a la utilización de un

cierto porcentaje de la muestra suministrada para el entrenamiento y el porcentaje restante para la prueba. La tercera opción es la utilización de una muestra de prueba distinta a la de entrenamiento que debe suministrarse al programa, la cual fue la que se utilizó para las pruebas de las muestras extraídas a partir del muestreo estratificado aleatorio que se mencionó previamente. Y la cuarta y última opción que brinda el programa mencionado es la conocida como validación cruzada.

3.2.2 Selección del *kernel*

Un paso fundamental en la aplicación de esta técnica es la selección del *kernel* o función núcleo, ya que es el que permite la transformación a un espacio de características sin necesidad de que se requiera conocer algún algoritmo explícito y, con esto, manejar datos no separables linealmente de manera mucho más sencilla. Existen diversos tipos de *kernel* que son comúnmente utilizados para diferentes tipos de estudio, entre los cuales se destacan el lineal, el polinomial, el sigmoidal y el RBF. Cada uno tiene una estructura particular y tiene asociado ciertos parámetros.

El *kernel* que se decidió utilizar para este estudio fue el RBF (*Radial Basis Function*), en español conocida como la función núcleo de función de base radial, la cual se define como

$$k(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \left\| \vec{x}_i - \vec{x}_j \right\|^2)$$

donde el parámetro γ debe ser mayor que cero, ya que controla el ancho del *kernel* y debe ser ajustado para un adecuado desempeño de la MVS. La decisión de la utilización de este *kernel* se fundamentó en el hecho de que diversos autores, entre los cuales se destaca Moro y Hurtado (2006) y Hsu *et al.* (2007), lo sugieren como una elección razonable para clasificación con MVS debido, fundamentalmente, a que ha demostrado buen desempeño en estudios previos. Además, dichos autores afirman que este *kernel* hace corresponder de manera no lineal ejemplos en un espacio de mayor dimensión y tiene menos hiperparámetros (C y γ) asociados a él que otros *kernels,* lo que hace menos complejo el modelo.

La selección del *kernel* lleva consigo la asignación de los valores de sus parámetros. Esta es una etapa trascendental en la aplicación de las MVS, por cuanto la obtención de resultados acertados dependerá, en gran parte, de la selección de un *kernel* adecuado y, por supuesto, de apropiados valores de los parámetros del mismo. Por lo tanto, se propusieron varias configuraciones, con la intención de determinar cuál de ellas arrojaba mejores resultados en cuanto a errores de clasificación. Las diferentes configuraciones propuestas se muestran en la tabla 5.

Configuración	C	γ
1	1	0,01
2	1	0,1
3	1	1,0
4	1	10,0
5	1	100,0
6	10	0,01
7	10	0,1
8	10	1,0
9	10	10,0
10	10	100,0
11	100	0,01
12	100	0,1
13	100	1,0
14	100	10,0
15	100	100,0
16	1000	0,01
17	1000	0,1
18	1000	1,0
19	1000	10,0
20	1000	100,0

Tabla 5: Configuraciones para las MVS

Uno de los parámetros cuyo valor es necesario determinar es el conocido como C, que se refiere al parámetro de penalidad para el error. En otras palabras, es el que va a permitir la holgura para tener cierto error de clasificación a cambio de tener una mejor generalización de los datos. Sin embargo, el valor de este parámetro varía de acuerdo a los datos que se desean clasificar; por lo tanto, se recomienda que se haga una búsqueda exhaustiva de un valor adecuado que permita clasificar correctamente la mayor cantidad de instancias desconocidas como sea posible.

Es importante tener en cuenta que con un valor muy elevado de C se tiene una alta penalización para puntos no separables y se tiende a tener muchos vectores soporte, lo que puede llevar al sobreajuste y, con esto, a la mala generalización. Por otra parte, un valor muy pequeño de C hace que el modelo sea muy rígido, lo que puede conducir a un subajuste. Tomando esto en cuenta, se decidió probar las MVS con configuraciones donde se variaba el valor de C entre 1 y 1000. Asimismo, el valor de γ se varió entre 0,01 y 100.

3.2.3 Entrenamiento

Con las muestras previamente seleccionadas, se procedió a realizar los distintos entrenamientos de las MVS. Para esto, se recurrió tanto a las muestras estratificadas

como a la muestra total de los datos. Es importante tener en cuenta que se realizaron entrenamientos con diferentes tamaños de muestras estratificadas (n = 198 y n = 1500) obtenidas previamente, con la intención de estudiar la capacidad de las MVS para clasificar correctamente cuando se entrenan con muestras bastante pequeñas y, por otro lado, cuando se entrenan con muestras relativamente grandes.

El clasificador utilizado para la fase de entrenamiento de las MVS planteadas fue el conocido como SMO por sus siglas en inglés *Sequential Minimal Optimization*, el cual implementa un algoritmo de optimización mínima secuencial. Este clasificador es el que ofrece *Weka,* el *software* seleccionado para esta tarea. De manera general, esta implementación lo que hace es reemplazar todos los valores faltantes y transformar los atributos nominales en binarios, normalizando por defecto todos los atributos. Este clasificador resuelve los problemas de multiclasificación utilizando clasificación por pares uno contra uno, con el cual se va realizando el modelo a partir de todas las posibles combinaciones de las diferentes clases.

Tal como se señaló anteriormente, el entrenamiento de una MVS requiere la solución de un problema muy grande de optimización de programación cuadrática. Platt (2003) afirma que SMO lo que hace es transformar ese problema grande en una serie de problemas de programación cuadrática más pequeños. Estos problemas más pequeños son resueltos analíticamente, lo que evita una optimización numérica desperdiciadora de tiempo. La cantidad de memoria requerida por SMO es lineal respecto al tamaño de la muestra de entrenamiento, lo cual le permite a SMO manejar conjuntos de datos muy grandes.

El clasificador SMO que brinda la herramienta *Weka* puede configurarse de acuerdo a los distintos parámetros asociados. Una de ellos se refiere al tipo de *kernel* que se desea utilizar, para los cuales proporciona dos posibilidades, el uso del *kernel* polinomial y el uso del *kernel* RBF. Además, la configuración del SMO es el que permite cambiar los parámetros asociados a los *kernels* como gamma (para el RBF), exponente (para el polinomial), tipo de filtro (si los datos serán normalizados o no), etc. Asimismo, en dicha configuración es donde se debe cambiar el valor del parámetros C.

3.2. 4 Pruebas y análisis de resultados

La última fase de clasificación con MVS consiste en realizar las respectivas pruebas al clasificador seleccionado para medir su desempeño. Como se mencionó anteriormente, inicialmente se utilizó la opción de suministrarle a la herramienta un conjunto de datos para la prueba compuesto de instancias diferentes a las incluidas en la muestra de entrenamiento, las cuales se extrajeron mediante muestreo aleatorio

estratificado. Y, posteriormente, se usó la opción de validación cruzada que brinda *Weka.*

Inicialmente, se probaron las 20 diferentes configuraciones propuestas para tres muestras de entrenamiento distintas compuestas de 1500 ejemplos. Luego, se calculó el promedio de instancias clasificadas correctamente para cada una de las configuraciones. Este promedio fue el que permitió determinar con cuál de dichas configuraciones se obtenían mejores resultados y, con esto, determinar la mejor configuración de MVS propuesta y poder estudiar su generalización para diferentes muestras de entrenamiento y de prueba. Los resultados obtenidos se encuentran plasmados en la tabla 6.

Posteriormente, se realizó el mismo procedimiento inicial pero para muestras de entrenamiento de 198 ejemplos. Es decir, se probaron las 20 configuraciones propuestas para tres muestras aleatorias estratificadas, se calculó el promedio de porcentaje de instancias clasificadas correctamente y se procedió a determinar con cuál de ellas se obtenía mejor clasificación. Los resultados obtenidos se pueden observar en la tabla 7.

Al realizar el entrenamiento de las MVS con 1500 ejemplos y tres de las muestras aleatorias extraídas con sus respectivas pruebas, la configuración que arroja mejores resultados, en cuanto a porcentaje de instancias clasificadas se refiere, es la denotada con el número 17. Al igual que cuando se entrenaron las MVS con 1500 ejemplos, cuando se entrenaron con solamente 198 ejemplos se obtuvo como mejor configuración la denotada con el número 17.Dicha configuración se refiere a un parámetro de regularización C igual a 1000 y un parámetro gamma de 0,1.

n = 1500			**Instancias clasificadas correctamente**			
Configuración	**C**	γ	**Muestra 1**	**Muestra 2**	**Muestra 3**	**Promedio**
1	1	0,01	97,3357%	97,3926%	97,3212%	97,3498%
2	1	0,1	97,3357%	97,3926%	97,3212%	97,3498%
3	1	1,0	97,8859%	97,4660%	97,6108%	97,6542%
4	1	10,0	98,1321%	97,4660%	97,8569%	97,8183%
5	1	100,0	97,5818%	97,4225%	97,3646%	97,4563%
6	10	0,01	97,3501%	97,3926%	97,3357%	97,3595%
7	10	0,1	97,9873%	97,5818%	97,6832%	97,7508%
8	10	1,0	98,6099%	97,8859%	98,1755%	98,2238%
9	10	10,0	98,3493%	98,0492%	98,3058%	98,2348%
10	10	100,0	97,6832%	97,6977%	97,7121%	97,6977%
11	100	0,01	98,0017%	97,5818%	97,7121%	97,7652%
12	100	0,1	98,6968%	97,8714%	98,3927%	98,3203%
13	100	1,0	98,8561%	98,3637%	98,3348%	98,5182%
14	100	10,0	98,1900%	98,2479%	98,3348%	98,2576%
15	100	100,0	97,7556%	97,6542%	97,7121%	97,7073%
16	1000	0,01	98,7113%	97,8569%	98,4506%	98,3396%
17	**1000**	**0,1**	**99,0153%**	**98,4769%**	**98,6968%**	**98,7297%**
18	1000	1,0	98,7547%	98,7257%	98,4217%	98,6340%
19	1000	10,0	98,2624%	98,2769%	98,3203%	98,2865%
20	1000	100,0	97,6108%	97,6542%	97,6832%	97,6494%

Tabla 6: Resultados de todas las configuraciones propuestas (n = 1500)

n = 198			Instancias clasificadas correctamente			
Configuración	**C**	γ	**Muestra 1**	**Muestra 2**	**Muestra 3**	**Promedio**
1	1	0,01	97,3687%	97,3444%	97,2713%	97,3281%
2	1	0,1	97,3687%	97,3444%	97,2713%	97,3281%
3	1	1,0	97,3687%	97,3444%	97,2713%	97,3281%
4	1	10,0	97,5271%	97,3444%	97,2713%	97,3809%
5	1	100,0	97,3687%	97,3444%	97,2713%	97,3281%
6	10	0,01	97,3687%	97,3566%	97,2713%	97,3322%
7	10	0,1	97,6611%	97,6855%	97,4297%	97,5921%
8	10	1,0	98,1240%	97,7586%	97,8682%	97,9169%
9	10	10,0	97,7707%	97,5149%	97,3444%	97,5433%
10	10	100,0	97,3687%	97,3200%	97,2713%	97,3200%
11	100	0,01	97,7342%	97,7098%	97,6124%	97,6855%
12	100	0,1	98,0266%	97,5880%	98,0631%	97,8926%
13	100	1,0	97,6855%	97,9169%	97,8682%	97,8235%
14	100	10,0	97,7707%	97,5758%	97,3444%	97,5636%
15	100	100,0	97,3687%	96,9546%	97,2713%	97,1982%
16	1000	0,01	97,9778%	97,5880%	98,0509%	97,8722%
17	**1000**	**0,1**	**97,6976%**	**98,0631%**	**98,0631%**	**97,9413%**
18	1000	1,0	97,6855%	97,7098%	97,8682%	97,7545%
19	1000	10,0	97,7707%	97,3200%	97,3444%	97,4784%
20	1000	100,0	97,3687%	96,9546%	97,2713%	97,1982%

Tabla 7: Resultados de todas las configuraciones propuestas (n = 198)

Luego de haber realizado las pruebas correspondientes a los diferentes tamaños de muestras de entrenamiento, se pudo proseguir a determinar, de manera general, la mejor configuración de MVS propuesta. Seguidamente, se realizó un estudio de generalización con diferentes muestras para la configuración establecida como la mejor, con el fin de estudiar la consistencia del modelo seleccionado y su desempeño respecto a la clasificación adecuada de instancias desconocidas.

De manera global, tanto para un tamaño de muestra de entrenamiento de 1500 como para un menor tamaño de 198, la configuración que arrojó mejores resultados fue la denotada con el número 17. Dicha configuración se refiere a valores de C y γ de 1000 y de 0,1, respectivamente. Además, se puede observar que lo que se pierde cuando se utilizan únicamente 198 ejemplos es menos de un error porcentual, en comparación de cuando se usan 1500 ejemplos para entrenar las máquinas. Esto reafirma la gran capacidad que tienen las máquinas de vectores soporte para la generalización, ya que con menos del 3% de la muestra original de los datos, las mismas logran clasificar muy bien las instancias desconocidas. Cabe destacar que la extracción de muestras a través del muestreo aleatorio estratificado contribuyó a lograr esto, ya que dicho tipo de muestreo permite reducir la varianza entre estrato y estrato y, con esto, se logran obtener muestras altamente representativas, sin importar su magnitud. Esto, a su vez, permite un adecuado entrenamiento, pruebas y, por lo tanto, se consigue buena generalización.

Una vez que se determinó que el mejor modelo se obtenía mediante la configuración 17 (C = 1000 y γ = 0,1), se pasó a estudiar su desempeño para diferentes

pruebas. Esta fase es la que permite medir la capacidad de clasificación para nuevas observaciones o, como comúnmente es conocido, la capacidad de generalización de las MVS, la cual es una de las características más atractivas de esta técnica. En este sentido, el objetivo que se busca es obtener un modelo que permita predecir correctamente la clase a la que pertenece cada instancia; es decir, dado una distribución de gastos de una determinada familia, especificar cómo es el comportamiento de consumo de acuerdo a las clases establecidas.

En la figura 2 se puede ver gráficamente el comportamiento del modelo seleccionado para las diferentes pruebas con un tamaño de muestra de entrenamiento de 1500 ejemplos, respecto al error absoluto medio y al error cuadrático medio. Este gráfico, en el cual se reflejan los valores de la tabla recientemente mencionada, permite estudiar la consistencia del modelo con respecto a estas medidas de desempeño, observándose muy poca variabilidad en el error absoluto medio y el error cuadrático medio. Asimismo, en la figura 3 se muestra el comportamiento del modelo respecto al porcentaje de instancias desconocidas clasificadas correctamente.

De la misma manera como se hizo para entrenamientos con 1500 ejemplos, para un tamaño muestral de entrenamiento de 198 ejemplos, se realizaron diversas pruebas, las cuales consistieron en entrenar con diferentes pruebas del mismo tamaño y probar su desempeño con muestras de diferentes tamaños, compuestas de instancias diferentes a las usadas para entrenar las MVS. Los valores obtenidos de las distintas pruebas fueron graficados y, posteriormente, analizados. En la figura 4 se muestra gráficamente los resultados obtenidos para las diferentes pruebas realizadas con un tamaño muestral de entrenamiento de 198, respecto a error absoluto medio y al error cuadrático medio. Mientras que en la figura 5 se muestra el comportamiento del modelo seleccionado, para muestras de entrenamiento de 198 ejemplos, respecto al porcentaje de instancias desconocidas clasificadas correctamente.

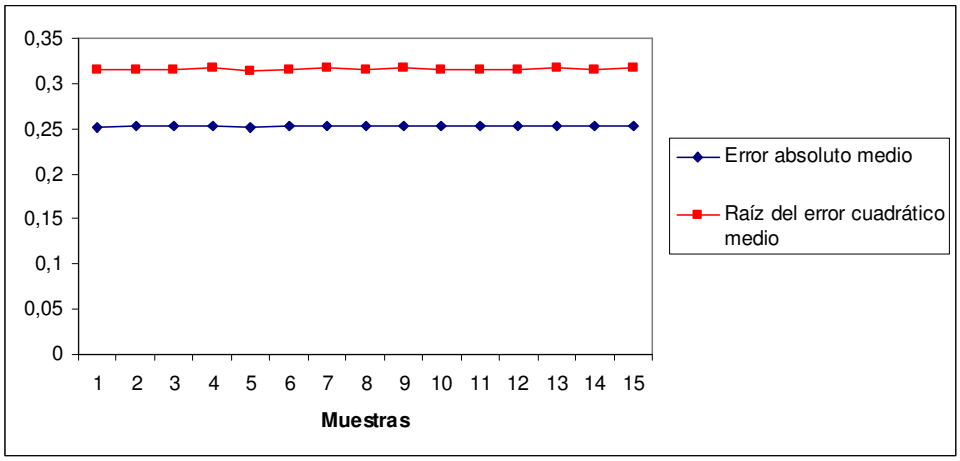

Figura 2: Errores para las pruebas de la mejor configuración (n = 1500)

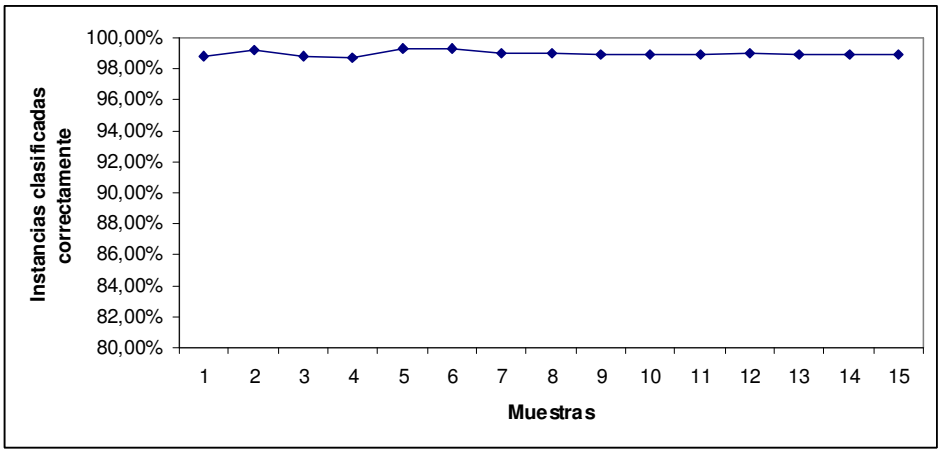

Figura 3: Resultados de clasificación para las pruebas de la mejor configuración ´ (n = 1500)

Para ambos tamaños muestrales de entrenamiento se observa muy poca variabilidad en las medidas utilizadas para evaluar el desempeño de la MVS, a saber: porcentaje de instancias clasificadas correctamente, error absoluto medio y raíz del error cuadrático medio. Esto permite afirmar que el modelo propuesto es consistente. Además, se observa que la configuración propuesta resulta muy buena generalizando pues el error en las muestras utilizadas es bastante pequeño y, asimismo, el porcentaje de instancias desconocidas clasificadas correctamente está cerca del 98% para todas las muestras probadas. Por lo tanto, se puede concluir que el modelo seleccionado arroja excelentes resultados respecto a consistencia y generalización, tal y como se esperó a priori.

Asimismo, se concluye que la metodología utilizada, la cual incorpora un adecuado preprocesamiento de los datos, el aprovechamiento de las bondades de la técnica estadística multivariante Análisis de Componentes Principales y la disposición de las cualidades de la técnica de aprendizaje automático Máquinas de Vectores Soporte, lograron producir un modelo de identificación de patrones de consumo robusto y consistente. Dicha metodología permitió encontrar que el 97,31% de los venezolanos se comportan, para el año de estudio, como consumidores de productos y servicios básicos. Es decir, casi todos los venezolanos dan prioridad a sus gastos básicos y luego a los gastos de lujo, lo cual se esperó a priori. Por otro lado, solamente el 0,63% de las familias venezolanas estudiadas dan prioridad a gastos de educación y servicios diversos. Y, por último, únicamente 0,81% dan prioridad a gastos de recreación y un 1,25% mostró que la mayoría de sus desembolsos estaban destinados a gastos financieros, tributarios y legales.

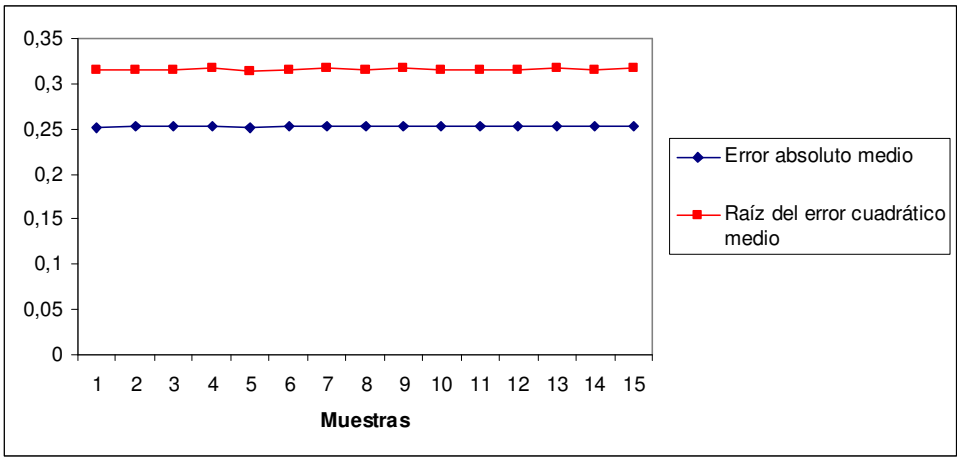

Figura 4: Errores para pruebas de la mejor configuración (n = 198)

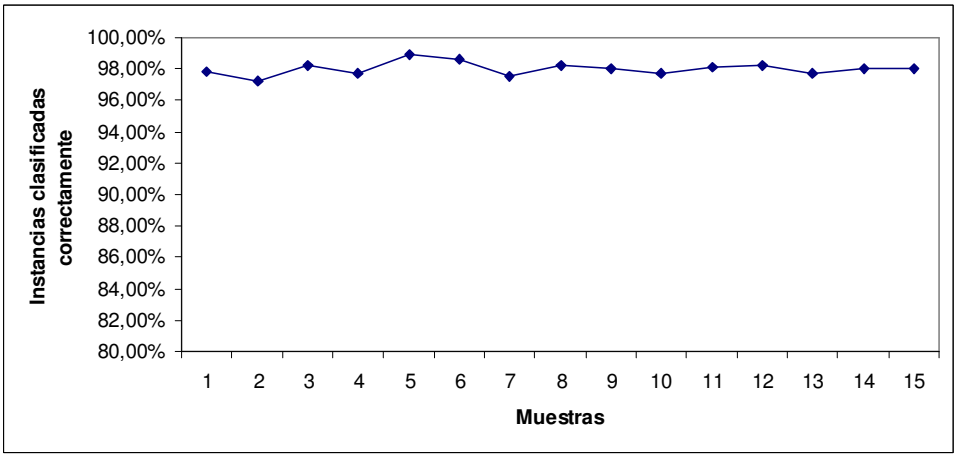

Figura 5: Resultados de clasificación para pruebas de la mejor configuración (n = 198)

CONCLUSIONES

- El preprocesamiento de los datos fue parte fundamental en la obtención de un modelo óptimo ya que resulta un paso útil para dar forma y coherencia a los datos originales, permitiendo que puedan ser manipulados para lograr el objetivo buscado.
- La aplicación de la técnica multivariante Análisis de Componentes Principales permitió definir cuatro maneras prevalecientes de consumo en el país para los años estudiados (1997-1998), a saber: 1) Gastos básicos; 2) Gastos de educación y servicios diversos, 3) Gastos de recreación; y 4) Gastos financieros, tributarios y legales.
- El Análisis de Componentes Principales, a través de sus bondades, facilitó la definición de la variable dependiente o de salida para las Máquinas de Vectores Soporte, lo que permitió distinguir las conductas de los venezolanos respecto a sus consumos.
- El uso de muestreo aleatorio estratificado para la selección de las muestras de entrenamiento permitió reducir la varianza entre estrato y estrato y, con esto, se logró

obtener muestras altamente representativas, sin importar su magnitud. Esto contribuyó adecuados entrenamientos, pruebas y, por lo tanto, una buena generalización.

- La metodología utilizado permitió identificar que para el año de estudio los venezolanos mostraron los siguientes patrones de consumo: 97,31% daban prioridad a gastos básicos; solamente un 0,63% mostraron inclinación hacia gastos de educación y servicios diversos; un 0,81% revelaron una propensión al gasto hacia gastos de recreación; y, por último, un 1,25% indicaron una tendencia a gastos financieros, tributarios y legales.
- La herramienta Máquinas de Vectores Soporte arrojó muy buenos resultados en la identificación de patrones de consumo de los venezolanos ya que proporcionó un modelo altamente consistente con una varianza del error absoluto medio casi nula.
- Las Máquinas de Vectores Soporte mostraron grandes capacidades de generalización ya que con el modelo conseguido, incluso cuando el entrenamiento se hizo con pocos ejemplos, el error absoluto para las distintas muestras se aproximaba a 0,25 y las instancias desconocidas clasificadas correctamente oscilaban cerca del 98%.
- Las MVS resultaron muy útiles en la identificación de patrones de consumo de los venezolanos, ya que a partir de la distribución del gasto de una determinada familia y de acuerdo a las clases previamente definidas, se pudo conocer su tipo de consumo.
- De manera general, se puede concluir que la metodología utilizada, la cual incorpora un adecuado preprocesamiento de los datos, el aprovechamiento de las bondades de la técnica estadística multivariante Análisis de Componentes Principales y la disposición de las cualidades de la técnica de aprendizaje automático Máquinas de Vectores Soporte, lograron producir un modelo de identificación de patrones de consumo robusto y consistente. Por lo tanto, se puede afirmar el gran beneficio que tienen los modelos híbridos obtenidos mediante la unión de diferentes áreas del conocimiento, como lo son la Estadística y la Ingeniería de Sistemas.

RECOMENDACIONES

Realizar estudios similares utilizando otros *kernels,* con la finalidad de tener parámetros de comparación en este ámbito de estudio respecto al desempeño de los modelos obtenidos mediante diferentes tipos de funciones núcleo.

Obtener los parámetros óptimos del *kernel* a través de un algoritmo o procedimiento formal, ya que de los mismos depende en gran parte el desempeño de las MVS para la clasificación o identificación de patrones de consumo.

Realizar estudios análogos al concluido pero utilizando los datos obtenidos de la III ENPF, la próxima IV ENPF y las subsiguientes a realizarse en el país, que permitan

observar los cambios en las prioridades de los venezolanos respectos a sus gastos, en el transcurso de los años.