PROYECTO DE GRADO

Presentado ante la ilustre UNIVERSIDAD DE LOS ANDES como requisito parcial para obtener el Título de INGENIERO DE SISTEMAS

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

Por

Br. Liz Jeannette Aranguren Pachano
Tutor: Gerardo A. Colmenares L., Ph. D.

Abril 2008

Identificación de patrones de consumo de los venezolanos mediante Máquinas de Vectores Soporte

Br. Liz J. Aranguren P.

Proyecto de Grado – Investigación de operaciones, 88 páginas

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. Además, entre las conclusiones se destaca que esta técnica, usada como método de clasificación no lineal, produce modelo consistente y mostró gran capacidad para generalizar, incluso cuando el entrenamiento se hizo a través de pocos ejemplos. Y, por último, 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

Este proyecto está dedicado con el amor y el afecto más profundo
a mi gran maestra y madre, Lizabeth Pachano,
a mi padre, Freddy Aranguren,
a mis hermanas Lizyvette y Lizddy.
y a mis sobrinos Kristabella, Kristoff y Melody,
quienes me inspiraron y motivaron a lograr esta meta.
Los Adoro.

Índice

Índice de Tablas

Índice de Figuras

Agradecimientos

En estas líneas deseo agradecer de corazón a todos los que influyeron en este estudio y, de alguna manera, permitieron que se hiciera posible. Antes que todo doy gracias a Dios Todopoderoso, por permitirme el alcance de esta meta bendiciéndome con vida y salud.

A mi madre Lizabeth Pachano, por su apoyo incondicional, motivación, orientación y ayuda. Por sus enseñanzas continuas y su confianza, por ser mi ejemplo a seguir.

Al profesor Gerardo Colmenares por haberme dado el privilegio de llevar a cabo esta grata investigación, por brindarme sus conocimientos y guiarme en el transcurso de su realización.

A mi padre Freddy Aranguren y mis hermanas Lizyvette y Lizddy, por motivarme día a día para lograr esta meta, por interesarse en el desarrollo de este estudio y creer en mí. Y a Zaur, Mireya y Edwin por quererme como familia y brindarme su cariño y solidaridad.

En especial a Francisco Justo, por haber estado a mi lado en los momentos más difíciles y más agradables vinculados con el progreso de esta investigación y por su ayuda constante.

A mi amiga y compañera Annjulie, quien compartió conmigo esta travesía, por sus palabras de aliento cuando más las necesité y porque juntas aprendimos que con paciencia y perseverancia lograríamos el sueño tan anhelado de convertirnos en Ingenieras de Sistemas.

A mis amigas de siempre, Ene, Daya, Amalia, Tita, Lucía, Virginia, Sori, Karo y Joha, por animarme continuamente y brindarme su valiosa amistad. Y a mis amigos y colegas cultivados en el transcurso de mi carrera, Justo, Amarú, Rafa, Maryury, Beatriz y Chynthia, por hacer más agradable esta difícil trayectoria.

Capítulo 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.

En el caso particular del presente estudio se hace referencia, primordialmente, a los patrones de consumo familiares, los cuales se pueden entender como estilos de comportamiento de grupos familiares, efecto de valores observados que siguen alguna regla o esquema de consumo. Para Sevilla (2006), el consumo es la síntesis de la actividad económica, pues representa la etapa del disfrute personal de bienes y servicios que produce una sociedad en conjunto.

Es indudable, entonces, la importancia de los patrones de consumo como indicadores para establecer las características socio-económicas nacionales. Dentro de dichas características es importante conocer, en relación a la familia, los recursos de los cuales dispone, lo que les debe a otros, los que genera de manera propia y la forma cómo los distribuye (Giordani, 2004). Estos elementos son considerados en las Encuestas Nacionales de Presupuestos Familiares, las cuales son definidas por Garnica (1996) como instrumentos para recabar información básica sobre condiciones de vida, ingreso y consumo de las familias. Además, dicha autora afirma que las encuestas mencionadas “arrojan un gran volumen de información

demográfica, sobre la situación y característica de la vivienda, educación, empleo, fuentes de ingreso y modalidades de consumo*"*(p. *55*).

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.

La selección de programas indicados para llevar a cabo la aplicación de las técnicas deseadas es un elemento importante cuando se quieren manejar grandes conjuntos de datos y emplear herramientas que tengan cálculos complejos inherentes a ellas. En el caso particular del estudio que se presenta, se recurrió programa *Statgraphics* en su versión de prueba 5.1 para ejecutar la técnica multivariante Análisis de Componentes Principales. Dicho programa permite realizar una gran variedad de análisis estadísticos a grandes volúmenes de datos; además, entre sus características atractivas se destaca su fácil manipulación y sus grandes capacidades gráficas. Por otro lado, el programa seleccionado para desarrollar la metodología de Máquinas de Vectores Soporte fue *Weka* en su versión 3.4.12, la cual es una

implementación en Java de varios algoritmos de clasificación, regresión, clustering, asociación y visualización. *Weka* tiene grandes ventajas entre las cuales se encuentra que es de fácil implementación debido a su interfaz gráfica y es de libre distribución y difusión.

Es importante destacar que, debido a la orientación que tienen las encuestas utilizadas para la realización del estudio, los patrones de consumo son interpretados a través de la manera cómo distribuyen su dinero las familias venezolanas. Es por esto que se habla de patrones de consumo familiares y no de patrones de consumo particulares de los venezolanos. Resulta fundamental tener presente a lo largo de todo el documento que las familias venezolanas, a las cuales se les estudió la propensión al gasto, son identificadas en las ENPF como hogares a los que se llegó a través de sus viviendas. Cada uno de dichos hogares está constituido por un grupo de venezolanos cuya relación puede estar o no fundamentada en nexos consanguíneos o legales. Los miembros de los hogares o "familias" comparten una misma vivienda principal y dependen de un fondo común para sus gastos, el cual está integrado por sus aportaciones y es utilizado para satisfacer necesidades tanto particulares como comunes de los mismos.

1.1 Antecedentes

Antes de comenzar el estudio referente a patrones de consumo familiares de los venezolanos fue necesario realizar una revisión bibliográfica con el fin de conocer algunas investigaciones que se han hecho en el mismo ámbito. Dichas investigaciones permiten, de alguna manera, fundamentar el estudio que se realizó debido a la relación directa entre ellas. Además, algunas de las que se mencionan a continuación sirvieron de guía y permitieron realizar comparaciones luego de haber identificado ciertos patrones de consumo familiares. La revisión de la literatura permitió encontrar algunas investigaciones relacionadas con el presente estudio, las cuales se citan a continuación:

Garnica (1996) condujo una investigación titulada Análisis de Componentes Principales en los presupuestos familiares en la ciudad de Mérida, planteándose como objetivos: representar, en forma gráfica, la información más importante de los gastos efectuados por las familias merideñas en 1986; describir el patrón de consumo de las familias observadas; y describir el papel que juega el ingreso en el presupuesto familiar. En las conclusiones se destaca que el gasto en el grupo de alimentos se mantiene más estable que el gasto en los

demás grupos de la canasta familiar. Sea cual sea el ingreso familiar, se cubren primero, las necesidades alimenticias y el resto del ingreso se destina a otros grupos. Además, se señala que los aspectos más importantes en los presupuestos familiares, en orden de prioridades, son: los gastos para el mantenimiento del hogar, los gastos en aspecto personal y los gastos en equipamiento de vivienda. La investigación realizada por Garnica resulta de gran interés por cuanto utiliza la técnica Análisis de Componentes Principales para destacar algunos rasgos del consumo familiar, lo cual demuestra que los estudios para identificar y describir patrones de consumo en nuestro país se hacían utilizando técnicas lineales. Además, se considera un aporte fundamental por cuanto brinda elementos básicos de información sobre las tendencias de consumo prevalecientes en las familias merideñas, a través de los datos obtenidos de una Encuesta de Presupuestos Familiares aplicada en 1986.

Anido (1998) realizó un trabajo con el propósito de estimar las elasticidades precio e ingreso de la demanda, para una muestra escogida de la ciudad de Mérida, durante el año 1986, a partir de la formulación de un sistema lineal del gasto. Para realizar el análisis multivariante de los datos, el investigador utilizó las técnicas Análisis de Componentes Principales y Análisis Factorial de Correspondencias Múltiples. Al final del estudio se verificó empíricamente que los bienes alimentarios, bienes en los que el consumidor merideño asigna una proporción del gasto total, se comportan (tal y como se esperó a priori) como bienes normales necesarios, al igual que los gastos correspondientes a vivienda y sus servicios. Con relación al resto de los bienes (vestido y calzado, hogar y diversos), la investigación permitió concluir que se comportan como bienes normales de lujo. Asimismo, se encontró que todos los bienes resultaron, en promedio para la muestra, inelásticos ante las variaciones en sus precios. Este estudio se convierte en una importante herramienta para el análisis dinámico de consumo, con las limitaciones que supone la representatividad de la muestra, la calidad de la información y sus características temporales, para la planeación económica tanto del gobierno como de los hogares y las empresas.

Márquez (2002) creó una base de datos con tecnología referencial en Microsoft Access, con la finalidad de aprovechar y tener mejor acceso a los datos obtenidos de las II Encuesta Nacional de Presupuestos Familiares aplicada en los años 1997 y 1998. Dicha base de datos, además de que permite realizar consultas sobre los datos de la encuesta, fue la que se utilizó para acceder a la información necesaria para la realización del presente estudio. Por otra parte,

Márquez utilizó el Análisis de Correspondencias Simples y Múltiples para describir la relación existente entre las distintas variables socio-económicas de carácter cualitativo. Asimismo, utilizó el Análisis de Componentes Principales para relacionar las variables socio-económicas cualitativas con las cuantitativas de los gastos del hogar. Entre las conclusiones obtenidas se destaca: primero, que Carabobo, Distrito Federal, Miranda, Anzoátegui, Zulia y Bolívar, son los estados en los cuales los hogares tienen mayores desembolsos; segundo, los hogares cuyo jefe del hogar tiene alto nivel educativo tienden a tener altos niveles de consumo; y, tercero, aquellos hogares con bajas características cualitativas (viviendas en condiciones poco favorables, bajo nivel educativo y/o desempleo del jefe del hogar, etc.) presentan bajos consumos en bienes de primera necesidad.

Anido, Orlandoni y Quintero (2005) realizaron un estudio sobre el consumo a partir de las Encuestas de Presupuestos Familiares (ENPF), entre los años 1967 y 2005, tomando como caso el de la ciudad de Mérida. El proceso de análisis se basó en modelos de sistema lineal de gasto (LES) mediante mínimos cuadrados ordinarios y sistemas de regresiones aparentemente no relacionadas, para calcular los coeficientes de elasticidad precio e ingreso para la ciudad de Mérida. Los resultados permitieron verificar empíricamente el carácter de bienes normales necesarios de alimentos (cuya asignación presupuestaria tanto al nivel nacional como en el caso estudiado es relativamente lo más importante) así como la escasa variabilidad en los gastos alimentarios ante cambios en las variables precio e ingreso de los consumidores. La importancia de este trabajo radica en el análisis de datos provenientes de las Encuestas de Presupuestos Familiares, en lapsos comprendidos en más de 30 años, permitiendo interpretar la evolución o cambios en los patrones de consumo de las familias merideñas, durante ese período.

Sosa (2006) llevó a cabo un proyecto de investigación que tuvo como finalidad conocer la estructura del gasto en el presupuesto familiar venezolano, utilizando los datos provenientes de la Encuesta de Presupuestos Familiares aplicada en el año 1988. Las técnicas multivariantes utilizadas para llevar a cabo el estudio fueron el Análisis de Componentes Principales (ACP) y el Análisis Factorial de Correspondencias Múltiples (AFCM). El ACP permitió determinar una relación opuesta entre alquiler de vivienda y sus servicios y equipos de transporte personal, alimentos, alimentación fuera del hogar y servicios de transportes; mientras que el AFCM proporcionó información sobre la relación existente entre familias que no poseen vivienda, en

contraposición con aquellas que sí poseen pero que están deficientes en cuanto a la construcción y a servicio de agua. Por otro lado, Sosa realizó un análisis de los ingresos de las regiones mencionadas a través del cual pudo determinar que la región Zuliana, opuesta a la región de los Andes y Guayana, presentó gran desigualdad en la distribución del ingreso. Un hallazgo importante de esta investigación fue la observación de las diferentes maneras cómo consumen los venezolanos de acuerdo a las distintas regiones estudiadas; por ejemplo, la región capital estaba completamente descrita por gastos en alquiler de vivienda y sus servicios y alimentación fuera del hogar, mientras que la región de los andes estaba caracterizada por gastos en vestido, alimentos y servicios de transporte. Sin embargo, los gastos que prevalecieron de manera general fueron alquiler de vivienda y sus servicios, alimentos y transporte. El estudio realizado por Sosa permitió tener conocimiento sobre cómo era el consumo de los venezolanos en las regiones investigadas y, además, ofreció información para hacer comparaciones con los resultados obtenidos.

1.2 Planteamiento del problema

La mayoría de las investigaciones relacionadas con la identificación de patrones de consumo han sido conducidas a través de métodos lineales tales como técnicas multivariantes, las cuales constituyen herramientas estadísticas que estudian el comportamiento de diversas variables al mismo tiempo. Sin embargo, los métodos multivariantes lineales poseen ciertas desventajas entre las cuales cabe señalar la presencia de variables inherentes al modelo, difíciles de cuantificar y observar, lo cual hace más compleja la interpretación de los datos y, por consiguiente, aumenta la posibilidad de error en los resultados. En atención a esto surgen los modelos no lineales, como una alternativa para sobrellevar la desventaja en el desempeño mostrado por la capacidad de pronóstico y clasificación de los métodos lineales; vale decir, los modelos lineales generalizados. Tal es el caso de los métodos multivariantes como el Análisis Discriminante, Análisis Factorial, Modelos de Ecuaciones Estructurales, que cada uno, de acuerdo a su particularidad algorítmica, permite agrupar y/o clasificar linealmente, ignorando el componente no lineal envuelto en el problema que se desea resolver.

En razón a lo antes expuesto, surgió la inquietud por realizar una investigación tendente a utilizar las Máquinas de Vectores Soporte (MVS) como técnica de clasificación no lineal para la identificación de patrones de consumo familiares de los venezolanos. Para esto,

se recurrió a los datos arrojados por la II Encuesta Nacional de Presupuestos Familiares, debido a la validez y confiabilidad que representan, al ser conducidas por equipos interdisciplinarios representantes de organismos reconocidos del país, tales como el Banco Central de Venezuela, la Universidad de Los Andes, el Instituto Nacional de Estadística y la Corporación Venezolana de Guayana.

1.3 Objetivos

1.3.1 Objetivo general

Identificar los patrones de consumo de los venezolanos y sus tendencias a partir de la II Encuesta Nacional de Presupuestos Familiares (ENPF), aplicada durante los años 1997-1998, utilizando las Máquinas de Vectores de Soporte (MVS) como técnica de clasificación no lineal.

1.3.2 Objetivos específicos

- Realizar un Análisis de Componentes Principales (ACP) para clasificar de manera lineal los datos provenientes de las ENPF y reducirlos a variables latentes.
- Utilizar la metodología MVS para realizar la clasificación e identificación de los patrones de consumo de las familias venezolanas en las zonas de influencia tratadas por las ENPF.
- Comparar los resultados obtenidos mediante los métodos ACP y MVS.

1.4 Justificación de la investigación

Luego de haber hecho una revisión de antecedentes se encontraron algunos significativos utilizando los resultados provenientes de las ENPF. Los trabajos encontrados se caracterizan por utilizar técnicas de análisis lineales y por estar delimitados a la ciudad de Mérida, entendiéndose este último hecho por la participación protagónica de la Universidad de Los Andes en las ENPF. En atención a lo antes expuesto, surgió el interés por realizar la investigación que se presenta, cuya importancia radica en la utilización de MVS, una técnica de clasificación no lineal que ha estado ganando popularidad en los últimos años debido a sus tantas características atractivas. Entre las características más importantes de las MVS se destaca la incorporación del principio de Minimización de Riesgo Estructural, que en algunos casos ha

demostrado ser superior al principio de Minimización de Riesgo Empírico, empleado por Redes Neuronales y otros métodos lineales convencionales. El principio de Minimización de Riesgo Estructural, el cual minimiza el límite superior del riesgo esperado, dota a las MVS con una mayor habilidad para generalizar, es decir, tener una alta capacidad de pronóstico para nuevas observaciones. Es interesante contrastar los resultados obtenidos mediante la aplicación de MVS a las ENPF con los trabajos antes citados, a fin de medir la capacidad predictiva de la técnica que se propone y cuantificar los niveles de sensibilidad a los cambios de los patrones debido a las nuevas observaciones.

Por otro lado, es importante destacar que, a diferencia de los antecedentes mencionados, el presente estudio analizó los datos para todo el territorio nacional. La intensión de estudiar los patrones de consumo familiar en las distintas regiones del país tratadas por las ENPF radicó en la determinación y comparación del comportamiento de consumo en cada región, debido fundamentalmente, a las características asociadas a factores internos y externos que establecen la inclinación al gasto de acuerdo a la cultura y sociedad en que se desenvuelve.

1.5 Organización del documento

En este documento se presentan ocho capítulos en total, incluyendo este capítulo introductorio. En el capítulo 2 se definen algunos términos relacionados con patrones de consumo entre los cuales se destaca el comportamiento del consumidor y la teoría económica. El capítulo 3 aporta las notas metodológicas sobre la II Encuesta de Presupuestos Familiares incluyendo, fundamentalmente, sus objetivos, ámbito geográfico, marco conceptual, variables investigadas e instrumentos básicos de muestreo. El capítulo 4 contiene fundamentos teóricos de la técnica multivariante Análisis de Componentes Principales que se utilizó con fines exploratorios. El capítulo 5 explica nociones básicas sobre las Máquinas de Vectores Soporte, exponiendo tanto sus fundamentos teóricos como matemáticos. El capítulo 6 se centra en la parte experimental inicial del estudio, es decir el preprocesamiento de los datos y la selección de las muestras. El capítulo 7 explica el procedimiento de la aplicación de los métodos ACP y MVS incluyendo los resultados obtenidos. Y, por último, en el capítulo 8 se presentan las conclusiones y recomendaciones del estudio realizado.

Capítulo 2

Patrones de consumo

Los continuos avances tecnológicos, la variedad de productos y servicios disponibles en el mercado y las diferentes necesidades del día a día son algunos de los factores que influyen en la manera cómo consumimos las personas. El consumo, visto en economía como la etapa final del proceso económico, es esencial para el desarrollo humano. Es por ello y mucho más que el estudio del análisis del comportamiento de cierta población respecto a su consumo resulta interesante dentro de muchos contextos.

Tal y como se ha mencionado anteriormente, los patrones de consumo familiares y personales se encuentran afectados por diversos factores a los que nos vemos expuestos los seres humanos a diario. Además, los patrones de consumo se construyen en cada grupo familiar de acuerdo a valoraciones, prioridades, percepciones e intereses de sus miembros. La manera cómo consume un grupo familiar se ve reflejado en la secuencia de elecciones y acciones de los miembros del hogar incluyendo la selección, compra, uso, mantenimiento y reparación de ciertos productos y servicios.

Para poder determinar los patrones de consumo es importante analizar aspectos referentes al comportamiento del consumidor por cuanto existe una relación directa entre ambos. Antes de profundizar en el tema del comportamiento del consumidor es importante señalar que el consumidor comúnmente es interpretado como el comprador final o el que compra para consumir, y que el consumo es entendido como la adquisición, uso y disposición de un bien o servicio, para satisfacer necesidades, deseos y preferencias.

Es importante destacar que existen diferentes tipos de consumo. Según González (2002) el consumo puede ser: a) básico, referido a aquello cuya falta parcial produce una condición material extremadamente precaria; b) asociado a una idea de bienestar humano más pleno, lo cual hace referencia a necesidades diferentes de las básicas; y, c) sobreconsumo,

referido a una expansión desproporcionadamente alta de las necesidades de consumo. Tal como se aprecia, el consumo guarda estrecha relación con el concepto de necesidad, sobre la cual existen diferentes teorías entre las que se destaca la teoría de Maslow, basada en la llamada jerarquía de necesidades humanas. Maslow (citado en Chiavenato, 2000) plantea que las necesidades humanas están distribuidas en una pirámide en cuya base están las necesidades más elementales y recurrentes (necesidades primarias: fisiológicas y de seguridad); mientras que en la cima se hallan las más sofisticadas y abstractas (necesidades secundarias: sociales, de autoestima y de autorrealización).

Schiffman y Lazar (2001), al analizar las necesidades en términos del comportamiento del consumidor, manifiestan que la mayoría de ellas "nunca se satisfacen por completo o permanentemente" (p.70). Además, estos autores establecen una relación entre las necesidades y las metas individuales, las cuales crecen y cambian sin cesar en respuesta a la condición típica del individuo, al ambiente, a sus interacciones con otras personas y a sus propias experiencias. Esto, a su vez, influye en el constante cambio de la manera de consumir.

2.1 Comportamiento del consumidor

El conocimiento sobre lo que es consumidor, consumo y necesidad permite definir el comportamiento del consumidor, sobre el cual se encuentran diversos conceptos, siendo uno de los más comunes el dado por Arellano, (2002) quien lo define como "aquella actividad interna o externa del individuo o grupo de individuos dirigida a la satisfacción de sus necesidades mediante la adquisición de bienes o servicios" (p.6). En este orden de ideas también es oportuno citar el concepto dado por Minard, Engel y Blackwell (2000, p. 6) quienes definen el comportamiento del consumidor como "las actividades que se efectúan al obtener, consumir y disponer de productos y servicios".

Es importante notar que en el primer concepto se destaca la influencia de los procesos mentales y emocionales de los individuos cuando actúan como consumidores. Por otro lado, el segundo concepto hace énfasis en tres procesos fundamentales, a saber: a) obtener, que hace referencia a las actividades que llevan a la compra de un producto, b) consumir, que busca las respuestas al cómo, dónde, cuándo y bajo qué circunstancias los consumidores utilizan los productos, y c) disponer, que incluye la forma cómo los consumidores se deshacen de los productos.

En los conceptos señalados se distingue la presencia de factores tanto internos como externos que influyen en el comportamiento del consumidor. En este sentido, autores como Kotler (1989) señalan los aspectos personales y psicológicos como factores internos y la cultura y aspectos sociales como factores externos. Arellano (2002) analiza estos aspectos como las variables del comportamiento, clasificándolas de la siguiente manera:

- **Variables de Influencia**: incluyen los aspectos biológicos, aspectos sociales, aspectos económicos y aspectos geográficos. Los aspectos biológicos hacen referencia a características importantes relacionadas con el consumidor tales como la edad, sexo, talla, etc. Los aspectos sociales están determinados por la cultura, las clases sociales, los grupos y las subculturas. Schiffman y Lazar (2001) definen la cultura como "la suma total de creencias aprendidas, valores y costumbres que sirven para dirigir el comportamiento de los miembros de una sociedad determinada, en su calidad de consumidores" (p.322), a la vez que interpretan la subcultura como "un grupo cultural distintivo que existe como un segmento identificable de una sociedad más amplia y más compleja" (p.346). Así, por ejemplo, un consumidor venezolano estaría influenciado por los valores y costumbres tanto del país, como los de una región particular (la andina, por ejemplo), a la vez que podría estar influenciado por su participación en grupos subculturales tales como los religiosos, los cuales han tenido un gran auge en nuestro país en los últimos años. Indudablemente que la clase social, determinada en término de estatus social a través de la riqueza, poder y prestigio, así como también los grupos sociales (conjunto de dos o más personas que interactúan para alcanzar metas ya sea individuales o colectivas) son también variables de influencia en el comportamiento del consumidor. En relación con los aspectos económicos se suele hablar de niveles de precios, ingresos, comerciales, publicidad e infraestructura. Sobre los aspectos geográficos es importante destacar la influencia de las condiciones físicas naturales del ambiente, tales como la temperatura, la altitud, la topografía, el clima y las comunicaciones, entre otros, en el comportamiento del consumidor.
- **Variables de procesamiento**: Estas hacen alusión a cómo el consumidor procesa los influjos provenientes de las variables de influencia, a través de sensaciones, percepción, motivación y actitudes. La motivación es un elemento dinámico que

cambia de manera constante al reaccionar ante las experiencias de la vida. Por otra parte, una actitud es "una predisposición aprendida para responder a un objeto o una clase de objetos de una manera uniformemente favorable o desfavorable" (London y Della, 1995, p.124). Con base en este concepto puede decirse, entonces, que las actitudes se caracterizan por ser aprendidas, por tener un objeto, por tener dirección e intensidad y por su tendencia a ser estables y generalizables.

- Variables de resultado: Estas son explicadas por los comportamientos de compra, de retención de publicidad, de lealtad a la marca. Stanton, Etzel y Walter (2004) ratifican que el comportamiento de compra de los consumidores finales se describe como un proceso de decisión influenciado, tal como se ha señalado, "por la información, las fuerzas sociales y de grupo, las fuerzas psicológicas y los factores situacionales" (p.128).

2.2 Principios de la Teoría Económica

Estudios y enfoques sobre el comportamiento del consumidor, en diferentes tiempos y espacios, ha dado origen a ciertos principios que emergen de la Teoría Económica. Estos son presentados por Arellano (2002) de la siguiente manera:

- Las necesidades y deseos de los consumidores son ilimitados, y por lo tanto, no se pueden satisfacer por completo. En vista de ello, tenderá a escoger aquella alternativa (bien o servicio) que maximice su satisfacción.
- La utilidad de un producto consiste en la satisfacción generada al consumidor. A medida que se adquieren más unidades de un producto, la utilidad total disminuye.
- Los consumidores son perfectamente racionales al tomar una decisión de compra, lo cual significa que sus decisiones se toman de forma independiente y que sus preferencias son constantes a lo largo del tiempo. (pp. 33-34)

Es importante el conocimiento de estos principios para la conducción de estudios tendentes a determinar patrones de comportamiento del consumidor, por cuanto se parte del principio fundamental de que los consumidores son perfectamente racionales, cuando se intenta aplicar una encuesta que busca clasificar las características de los consumidores. Sin

embargo, la constancia de sus preferencias a lo largo del tiempo podría ser tema de análisis en cualquier investigación.

2.3 Interdisciplinariedad de los estudios sobre el comportamiento del consumidor

Los estudios sobre el comportamiento del consumidor están enfocados hacia la forma cómo los individuos toman decisiones para gastar sus recursos disponibles (tiempo, dinero, esfuerzo) en artículos relacionados con el consumo (bienes y servicios). Dada la importancia de estos comportamientos en los procesos socioeconómicos de las sociedades, cada día es mayor el interés en el tema. Es así como estos estudios, en palabras de Solomon (1997), adquieren la característica de ser interdisciplinarios, por cuanto "está compuesto por investigadores de diferentes campos que comparten un interés sobre las interacciones de las personas en el mercado" (p.45).

En este sentido, en trabajos sobre el comportamiento del consumidor es común encontrar, tal como lo expresan Schiffman y Lazar (2001), aportes de la sociología (características de los grupos), de la psicología social (estudios de las formas en que se desenvuelve el individuo en los grupos), de la antropología (influencia de la sociedad sobre el individuo) y de la economía (estudios sobre beneficios en la adquisición de bienes y servicios) por citar algunas ramas del saber. Sumado a esto, aún bajo el entendido de que toda ciencia tiene su propio método, es innegable los aportes de la matemática, la estadística y las ingenierías (particularmente la Ingeniería de Sistemas), en la búsqueda y aplicación de técnicas de procesamiento y análisis de los datos obtenidos en estudios sobre el comportamiento del consumidor, a fin de identificar y describir patrones de comportamiento, en poblaciones y muestras particulares de investigación.

Capítulo 3

Encuestas Nacionales de Presupuestos Familiares

Inicialmente, las encuestas de presupuestos familiares realizadas en el país fueron dirigidas a regiones específicas. En el año 1939, fue cuando se realizó la primera encuesta de presupuestos familiares, dirigida por el despacho de Fomento y aplicada sólo en la ciudad de Caracas. Luego, en el año 1967 se aplicó una encuesta de presupuestos familiares en la cuidad de Mérida, dirigida por el Instituto de Investigaciones Económicas y Sociales (IIES), de la entonces Facultad de Economía de la Universidad de Los Andes (ULA). Más tarde, durante los años 1988-1989 se llevó a cabo la denominada I Encuesta de Presupuestos Familiares por diversos organismos del país entre los cuales se destacan la ULA y el Banco Central de Venezuela (BCV). Posteriormente, en el año 1997, se comienza a aplicar la II Encuesta Nacional de Presupuestos Familiares (ENPF), la cual fue la fuente de información para el estudio realizado.

Una encuesta sobre presupuestos familiares es "una investigación estadística por muestreo que se realiza a los hogares con el fin de obtener información sobre sus ingresos, egresos, características de las viviendas que habitan, composición del hogar y otras variables económicas y sociales de los miembros que lo integran" (Meza, 2001, p. 9). Teniendo en cuenta que presupuesto es un concepto macroeconómico que demuestra la compensación hecha cuando se intercambia un bien por otro, se puede interpretar un presupuesto familiar como la estimación de gastos e ingresos en un período de tiempo específico para una familia determinada.

El Banco Central de Venezuela (BCV) y la Oficina Central de Estadística e Informática (OCEI), junto con la Corporación Venezolana de Guayana (CVG) y la Universidad de Los Andes (ULA) condujeron el desarrollo de la II Encuesta Nacional de Presupuestos Familiares a objeto de analizar los cambios en los patrones de consumo de los hogares ocurridos desde

1989, año en el cual se aplicó la primera Encuesta de Presupuestos Familiares. Entre sus finalidades se señalan: a) actualizar el Índice de Precios al Consumidor (IPC) en cuanto a la composición de la canasta y estructura de ponderaciones de los bienes y servicios que la componen; b) elaborar las principales cuentas del sector hogares, según los requerimientos del Sistema de Cuentas Nacionales (SCN); y, c) disponer de un conjunto de indicadores sociales para el estudio de las condiciones de vida de los hogares.

La II Encuesta Nacional de Presupuestos Familiares requirió de 15 meses para la recolección de los datos, comprendidos entre Enero de 1997 y Marzo de 1998. Desde el punto de vista geográfico, la investigación abarcó todo el territorio nacional. La muestra seleccionada fue de aproximadamente 9.900 hogares para una población estimada de 22.777.153 habitantes. Los dominios de estudio contemplados fueron: primero, el área metropolitana de Caracas; segundo, las ciudades principales del país (Maracay, Valencia, Barquisimeto, Maracaibo, Puerto La Cruz y Ciudad Guayana); tercero, las ciudades satélites de Caracas (Los Teques, Guarenas, Guatire, Ocumare del Tuy, etc.); cuarto, las ciudades medianas y grandes; y, por último, las ciudades de 25.000 o menos habitantes.

Para efectos del entendimiento y desarrollo de la II ENPF es necesario tener en cuenta algunos conceptos fundamentales. Dichos conceptos ayudarán a tener una visión más clara del enfoque que tiene la encuesta menciona. Una de las palabras más comunes que se maneja en los datos de la II ENPF es hogar. Para Meza (2001) un hogar está constituido por un conjunto de personas que comparten o dependen de una bolsa o fondo común para sus gastos; además, ese fondo está conformado por las aportaciones de los miembros y tiene como fin satisfacer las necesidades de sus miembros, tanto particulares como comunes.

Otro concepto fundamental es el de vivienda. Una vivienda es toda estructura hecha o no para vivir, pero utilizada con ese fin, en la cual puede habitar más de un hogar (Meza, 2001). Para efectos del estudio que se realizó se consideraron los habitantes de una vivienda como familias, debido al enfoque que se le dio donde los venezolanos fueron estudiados como familias venezolanas que habitan en un mismo espacio geográfico. Lo anterior se fundamenta en el hecho de que se pretendió observar, entre otras cosas, diferentes patrones de consumos en diferentes entidades del país.

3.1 Períodos de referencia de las II ENPF

Los periodos de referencia son "los períodos de tiempo, dentro de los cuales va a ser considerada la realización de determinados pagos o la percepción de ciertos ingresos" (Meza, 2001, p. 15). Trabajar con ciertos períodos de referencia fue necesario debido a que tanto los gastos como los ingresos son variables que no se efectúan o perciben regularmente en un cierto período.

Para que la aplicación de la II ENPF resultara válida y eficiente fue necesario definir los períodos de referencia de acuerdo a la capacidad de los encuestados para recordar, la importancia de los distintos gastos y la frecuencia con que se realizan los mismos. Los gastos que son más frecuentes y/o los que son más difíciles de recordar debido a sus bajos montos, se clasificaron como los períodos de menor referencia. Mientras que aquellos gastos más costosos y, por lo tanto, más fáciles de recordar se clasificaron dentro de períodos de referencia más altos.

De acuerdo a lo mencionado anteriormente, se consideraron gastos semanales todos aquellos que se realizan con alta frecuencia de consumo y/o con precios bajos. Entre los gastos semanales están todos los asociados a alimentos y bebidas, artículos de cuidado personal, periódicos, pasajes de transporte público, etc. Por otro lado, dentro de los gastos mensuales se incluyen pagos en alquileres de viviendas, mensualidades escolares o de estacionamientos, servicios del hogar como luz, agua, teléfono, etc. Asimismo, aquellos egresos en servicios y/o productos menos frecuentes que los anuales pero más frecuentes que los mensuales como gastos en muebles, equipos del hogar, servicios de salud y mantenimiento de la vivienda fueron clasificados como gastos trimestrales. Y, por último, los gastos menos comunes y/o con un valor elevado que facilita al encuestado recordarlos, como compra de vehículos, seguros, pagos de impuestos y remodelaciones de la vivienda, se etiquetaron como gastos anuales.

Los períodos de referencia para los ingresos fueron: mensual y anual. Los ingresos que fueron etiquetados como mensuales fueron aquellos que se tienden a recibir con más frecuencia como sueldos, salarios, honorarios profesionales, becas, etc. Los ingresos anuales incluyen beneficios que se tienden a obtener con poca frecuencia a través de utilidades, herencias en dinero, premios de loterías, etc.

3.2 Variables investigadas por la II ENPF

La II Encuesta Nacional de Presupuestos Familiares contempló fundamentalmente el estudio de tres variables para su estudio. Dichas variables son: los gastos del hogar, los ingresos del hogar y otras variables socio-demográficas. Los gastos están referidos a aquellos bienes y servicios que han sido adquiridos por algún miembro de la familia, sin importar quién lo haya pagado e independientemente de en qué período haya sido consumido. Los ingresos del hogar están referidos a todos los ingresos que han sido recibidos por sus miembros sin importar su origen ni su naturaleza. Por otra parte, algunas de las demás variables socio-demográficas son características de las viviendas (tipo, apariencia, disponibilidad de los servicios básicos, etc.), características de los miembros del hogar (nombre, sexo, edad, nivel de educación alcanzado, etc.) y dotación y equipamiento del hogar.

3.3 Instrumentos básicos de muestreo de la II ENPF

Los instrumentos básicos que se utilizaron en las II ENPF para la recolección de información fueron ocho cuestionarios llamados EPF. Cada EPF constó de diversas preguntas relacionadas con el nombre de cada una de ellas. Las ochos EPF fueron: 1) Unidades básicas de muestreo; 2) Características de las viviendas; 3) Características generales del hogar; 4) Gastos diarios del hogar; 5) Gastos diarios personales; 6) Gastos mensuales; 7) Gastos trimestrales y anuales; y 8) Ocupación e ingresos.

Las EPF que resultaron fundamentales para que llevar a cabo el estudio fueron aquellas donde se especificaban todos los tipos de gastos e ingresos, la entidad a la que pertenecía cada una de las viviendas y el número de miembros para cada hogar. Esto anterior se debe a que las variables mencionadas fueron las que se decidieron tomar en cuenta para la investigación sobre el consumo de los venezolanos.

Capítulo 4

Análisis de Componentes Principales

El Análisis de Componentes Principales (ACP) es una técnica estadística ubicada dentro del campo de los métodos multivariados, es decir métodos de análisis de datos de variables múltiples. Dicha técnica multivariante fue iniciada por Pearson en el año 1901 y, posteriormente, desarrollada por Hotelling en el año 1933. Johnson (2000) afirma que los métodos multivariados son extraordinariamente útiles para ayudar a los investigadores a hacer que tengan sentido conjuntos grandes, complicados y complejos de datos que constan de una gran cantidad de variables medidas en números grandes de unidades experimentales. Una unidad experimental es cualquier objeto o concepto que se puede medir o evaluar de alguna manera. Para el caso particular del estudio realizado, las unidades experimentales fueron las familias venezolanas, las cuales se estudiaron a través de las II Encuestas Nacionales de Presupuestos Familiares con el fin de identificar sus patrones de consumo.

Los métodos multivariados pueden ser utilizados desde diferentes enfoques entre las cuales se encuentran la simplificación de la estructura de datos, la clasificación y el análisis de interdependencia. Es importante tomar en cuenta que existen diversos tipos de métodos multivariados que pueden ser utilizados de acuerdo al interés particular de cada estudio y a los tipos de datos con los que se estén trabajando. Por lo tanto, cada situación requiere una evaluación particular para utilizar el método de análisis multivariado más adecuado, que permita extraer la máxima información posible del conjunto de datos, pero que a su vez garantice la validez de su aplicabilidad (Pla, 1986). Además, la mencionada autora afirma que el ACP deberá ser aplicado cuando se desee conocer la relación entre los elementos de una población y se sospeche que en dicha relación influye de manera desconocida un conjunto de variables o propiedades de los elementos.

El ACP, al igual que otros métodos multivariados, permite detectar interdependencia entre variables y también entre individuos. Una de sus grandes ventajas está en el hecho de que permite la estructuración de un conjunto de datos multivariados, sin necesidad de conocer su distribución de probabilidades. El procedimiento básico del ACP consiste en transformar el conjunto de variables originales de los datos a un nuevo conjunto menor de variables denominadas componentes principales. Es fundamental destacar que las variables del conjunto original de los datos se presuponen correlacionadas ya que miden información común mientras que las nuevas variables no están correlacionadas, con lo que se evita la redundancia de información. Además, las nuevas variables resultan de combinaciones lineales de las originales y se construyen de acuerdo al orden de importancia en cuanto a la variabilidad total que recogen de la muestra. Se hace notorio entonces que la varianza o variabilidad juega un papel fundamental en el ACP ya que a mayor varianza total presente en los componentes principales se mantendrá mayor información de las variables originales.

Anteriormente, estudios sobre patrones de consumo han sido realizados utilizando el ACP. Sin embargo, han sido delimitados sólo a la ciudad de Mérida. Cabe destacar que, para el caso particular de este estudió, el ACP fue utilizado sólo como un primer paso con fines exploratorios. Esta técnica permitió reducir la dimensionalidad de los datos agrupando las unidades experimentales (los venezolanos) en subgrupos de acuerdo a sus tendencias de propensión al gasto. El carácter cuantitativo de los datos fue una de las razones por las cuales se decidió utilizar el ACP como técnica exploratoria para reducir el número de variables a un menor número de variables latentes, las cuales serían utilizadas para la identificación de los patrones de consumo mediante las MVS.

4.1 Fases de un Análisis de Componentes Principales

4.1.1 Cálculo de los componentes

La base del análisis estadístico multivariado y, por lo tanto, la del análisis de componentes principales es el enfoque matricial y matemático. El cálculo de los componentes se puede hacer sobre una matriz de varianza-covarianza de las muestras o sobre una matriz de correlación. El mejor tipo de matriz suele depender de las variables que se desean medir. Sin embargo, no tiene ningún sentido aplicar un ACP a un conjunto de variables que no tengan altas correlaciones entre sí ya que esto implicaría que no existe información redundante y, por

lo tanto, no se puede reducir la dimensionalidad de los datos sin perder mucha información. Es por esto que se recomienda que antes de proceder a realizar el cálculo de los componentes se haga un breve análisis de la matriz de correlaciones de las variables para verificar que tenga sentido aplicar un ACP. Luego de verificadas las correlaciones de las variables se puede comenzar con el principal paso de esta técnica que es el cálculo de los componentes.

4.1.1.1 Matriz de varianza-covarianza

Cuando se desea calcular los componentes principales a través de la matriz de varianzas-covarianzas es importante tener en cuenta que este método puede arrojar resultados no muy confiables cuando existen grandes diferencias entre las varianzas de las variables. Este caso se presenta cuando existen, por ejemplo, variables medidas en kilómetros y otra(s) en metros, ya que ésto llevaría a dar mayor importancia a aquellas cuyas magnitudes sean considerablemente mayores. En otras palabras, se recomienda el uso de este método sólo cuando las varianzas de las variables a estudiar son similares y las mismas son medidas en una misma escala.

Para el cálculo de los componentes principales se tiene un conjunto original de p variables $(x_1, x_2, ..., x_p)$ sobre un grupo de objetos o individuos, con las cuales se obtendrá un nuevo conjunto de variables $(y_1, y_2, ..., y_p)$ resultantes de combinaciones lineales de las primeras y que estarán incorrelacionadas entre sí. La varianza de y_1 será mayor que la de y_2, la de y_2 mayor que la de y_3 y así sucesivamente; además, juntas explicarán el 100% de la variabilidad total de los datos originales. Cada nueva variable se puede expresar de la siguiente manera:

$$y_j = a_{j1}x_1 + a_{j2}x_2 + ... + a_{jp}x_p = \mathrm{a}'_j \, \mathrm{x} \qquad (4.1)$$

donde $\mathrm{a}'_j = (a_{1j}, a_{2j}, ..., a_{pj})$ es un vector de constantes y $\mathrm{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$.

Tomando en cuenta que se desea maximizar la varianza, pues a mayor varianza contenida en las nuevas variables se retiene mayor información de las variables originales, resulta lógico que se hace deseable tener valores altos de los coeficientes a_{ij}. De este modo, el primer componente se calcula seleccionando a_1 para que y_1 tenga la mayor varianza posible, el segundo componente con a_2 para que y_2 explique la mayor variabilidad posible que no

haya sido explicada por y_1 y, a su vez, que y_1 y y_2 no se correlacionen entre sí, y así para los p componentes. Para esto, es necesario mantener la ortogonalidad de la transformación lo cual se logra haciendo que el módulo de a'_j sea igual 1, es decir que $a'_j a_j = 1$; entonces, $y_1, y_2, ..., y_p$ deben cumplir con esta restricción.

El método de los multiplicadores de Lagrange juega un papel fundamental en el proceso de extracción de factores ya que es el indicado para resolver problemas donde se busca maximizar una función de varias variables que está sujeta a ciertas restricciones. Aplicando propiedades de multiplicadores de Lagrange se obtendrá una matriz de covarianzas S de orden p, cuya diagonal principal estará formada por los autovalores $\lambda_1, \lambda_2, ..., \lambda_p$ correspondientes a $a_1, a_2, ..., a_p$.

Para la búsqueda del primer factor se debe maximizar la función $a'_1 \sum a_1$, la cual maximiza la varianza de y_1, sujeta a una cierta restricción ($a'_1 a_1 = 1$), donde lo que se busca es encontrar el mejor vector a_1 que brinde la combinación lineal óptima que se desea. El mayor autovalor λ_1 es el que corresponde al autovector a_1, el cual arroja la combinación lineal de las variables originales que explican la mayor variabilidad entre ellas. Luego de tener $a'_1 = (a_{11}, a_{12}, ..., a_{1p})$ se puede calcular el primer factor de la siguiente manera:

$$y_1 = a'_1 x = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p \qquad (4.2)$$

Para encontrar el segundo componente se procede se manera similar que para el primero, a diferencia de que para éste se debe tomar en cuenta que él debe estar incorrelacionado con y_1, lo que añade otra restricción al problema que se desea resolver. Tomando en cuenta que se debe cumplir que $Cov(y_2, y_1) = 0$, donde $Cov(y_2, y_1) = a'_2 \sum a_1$, lo que se busca es maximizar la función $a'_2 \sum a_2$ que maximiza la varianza de y_2 sujeta a las restricciones $a'_2 a_2 = 1$ y $a'_2 a_1 = 0$. Usando propiedades de los multiplicadores de Lagrange y el mismo razonamiento, se elige λ_2 como el segundo mayor autovalor de la matriz $\sum$, cuyo autovector asociado a_2 permite encontrar a $a'_2 = (a_{21}, a_{22}, ..., a_{2p})$ y, con esto, el segundo componente:

$$y_2 = a'_2 x = a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p \qquad (4.3)$$

Utilizando procedimientos similares a los usados para encontrar los primeros dos componentes se procede a calcular los p componentes. Luego de obtenidos dichos componentes, los mismos se expresan como $\mathbf{y} = Ax$, en donde A es la matriz de los autovectores y x el vector formado por las variables originales. Así se obtiene la matriz de componentes de la siguiente manera:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \qquad (4.4)$$

4.1.1.2. Matriz de correlación

Cuando se decide utilizar la matriz de correlaciones para el cálculo de los componentes principales es necesario tener en cuenta que los cálculos se realizan sobre las variables originales estandarizadas, es decir, variables con media 0 y varianza 1. Según Johnson (2000), cuando las variables no se presentan con fundamentos iguales, es necesario aplicar el ACP a los datos estandarizados, lo que se hace es calcular los eigenvalores y eigenvectores a partir de la matriz de correlaciones. Dichos eigenvalores y eigenvectores son también conocidos como valores propios y vectores propios, respectivamente. La suma de todos los autovalores será p (número de variables estudiadas) y la proporción de variabilidad recogida por cada componente será λ_j / p.

Una de las características básicas de la matriz de correlaciones es que su diagonal principal está formada de unos; esto se debe a que las nuevas variables estandarizadas poseen varianza unitaria. Es importante destacar que el uso de la matriz de correlaciones implica una ponderación de las variables originales, otorgándole a cada una la misma importancia, independientemente de los valores relativos de sus varianzas.

El procedimiento para el cálculo de los componentes se hace bajo el mismo razonamiento que cuando se calculan a partir de la matriz de varianza-covarianza. Además, es importante tener en cuenta que la matriz de correlaciones (R) se relaciona con la matriz de varianzas-covarianzas (S) mediante la expresión $R = D^{-1/2} S D^{-1/2}$, donde $D^{-1/2} = Diag(1/s_i)$.

4.1.2 Selección del número de componentes principales

Luego de haber calculado los p componentes resultantes de combinaciones lineales de las p variables originales, se debe seleccionar un número de factores a los cuales se les denominará componentes principales. Sin embargo, la cantidad de factores que se deben retener es completamente subjetivo, dependiente del investigador, ya que no existen criterios formales que determinen el número preciso que se debe retener. El criterio que debe permanecer latente es que se desea mantener la mayor variabilidad de los datos con la menor cantidad de componentes principales posible. Por lo tanto, el número de factores a seleccionar dependerá de la variabilidad acumulada que se considere suficiente.

Cabe destacar que existen diversos criterios gráficos que se pueden utilizar como ayuda para seleccionar un número acertado de componentes principales. Díaz (2002) menciona que uno de ellos consiste en ubicar puntos en un diagrama cartesiano cuyas coordenadas son los componentes principales o factores y los valores propios. La idea es buscar una especie de "codo", es decir un punto a partir del cual se pueda trazar una línea recta de pendiente pequeña, y el número de componentes estará dado por los puntos ubicados arriba de tal línea. En el ejemplo de la figura 4.1 los tres primeros factores son los candidatos a escoger como componentes principales.

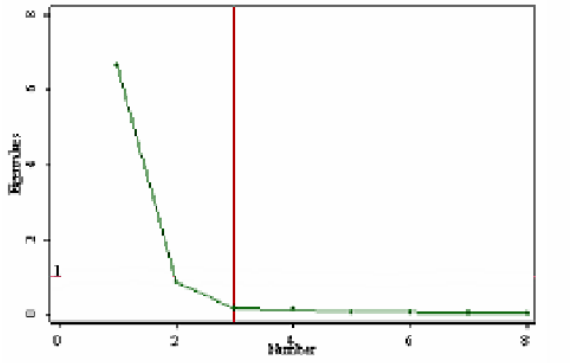


Figura 4.1: Ejemplo de selección gráfica del número de componentes principales

Otro método gráfico sugerido por Díaz (2002) es muy similar al anterior, la diferencia radica en que lo que se debe graficar son los componentes en orden decreciente (eje x) contra el porcentaje de variación explicado por cada uno de ellos (eje y). Sin embargo, la idea del "codo" es la misma, entonces lo que se busca es el punto a partir del cual los porcentajes de variación tienden a formar una línea recta de poca pendiente como se hace con los valores

propios para el método anterior. Así, los puntos a escoger serán los que se ubiquen por encima de dicho punto.

Cuando se han calculado los componentes a partir de la matriz de correlaciones se puede utilizar el criterio de seleccionar aquellos componentes cuyos autovalores sean mayores o iguales a 1. Johnson (2000) afirma que esto se debe a que se tiene la creencia de que si un componente principal no puede explicar más variación que una sola variable por sí misma, entonces es probable que no sea importante. Para esto es primordial tener claro que cuando se trabaja con datos estandarizados (matriz de correlación) cada variable original tendrá media 0 y varianza igual a 1. Este criterio pierde sentido cuando se trabaja con la matriz de varianzas-covarianzas.

4.1.3 Análisis de la matriz factorial

Una vez que se hayan seleccionado los componentes principales, los mismos se deben representar en forma matricial. De esta manera, la matriz quedará conformada por los coeficientes factoriales de las variables; es decir, los coeficientes de correlación entre las variables y los componentes principales. Dicha matriz tendrá tantas filas como variables y tantas columnas como componentes principales. Para esto, es importante tomar en cuenta los tanto la magnitud de los coeficientes como su signo.

4.1.4 Interpretación de los componentes principales

La interpretación de los componentes principales es un paso fundamental en la aplicación de la técnica pues el éxito de los resultados que se expresen dependerá, en gran parte, de que los componentes principales hayan sido interpretados correctamente. No obstante, esta interpretación es completamente subjetiva y resulta bastante compleja e incluso, en muchas ocasiones, imposible. Para poder llevar a cabo un buen análisis o interpretación de los componentes principales el analista o investigador debe tener un alto conocimiento sobre los datos que se manejan y las variables utilizadas para estudiar.

Un factor se hace más fácil de interpretar cuando coeficientes factoriales están próximos a 1 o tienen valor elevado. Además, es ventajoso cuando una variable tiene coeficientes elevados en un solo o pocos factores. Por otro parte, la interpretación de los componentes principales abarca el análisis tanto de correlaciones positivas como negativas pues éstas últimas también pueden aportar bastante información.

4.1.5 Cálculo de las puntuaciones factoriales

La última de las fases del ACP consiste en calcular las puntuaciones factoriales. Éste cálculo permite, entre otras cosas, tratar las puntuaciones en los *k* componentes principales como *k* nuevas variables. Con esto se justifica que el análisis de componentes principales se hace muy útil cuando se desea reducir un gran número de variables a un número menor de variables latentes. La manera para calcular las puntuaciones de los componentes principales es a través de las siguientes expresiones:

$$CP_1 = coef(f_1c_1)\cdot\frac{valor\_\mathrm{var}1-\overline{X}_{\mathrm{var}1}}{s_{\mathrm{var}1}} + coef(f_2c_1)\cdot\frac{valor\_\mathrm{var}2-\overline{X}_{\mathrm{var}2}}{s_{\mathrm{var}2}} + \ldots + coef(f_nc_1)\cdot\frac{valor\_\mathrm{var}n-\overline{X}_{\mathrm{var}n}}{s_{\mathrm{var}n}}$$

$$CP_2 = coef(f_1c_2)\cdot\frac{valor\_\mathrm{var}1-\overline{X}_{\mathrm{var}1}}{s_{\mathrm{var}1}} + coef(f_2c_2)\cdot\frac{valor\_\mathrm{var}2-\overline{X}_{\mathrm{var}2}}{s_{\mathrm{var}2}} + \ldots + coef(f_nc_2)\cdot\frac{valor\_\mathrm{var}n-\overline{X}_{\mathrm{var}n}}{s_{\mathrm{var}n}}$$

$$\vdots$$

$$CP_k = coef(f_1c_k)\cdot\frac{valor\_\mathrm{var}1-\overline{X}_{\mathrm{var}1}}{s_{\mathrm{var}1}} + coef(f_2c_k)\cdot\frac{valor\_\mathrm{var}2-\overline{X}_{\mathrm{var}2}}{s_{\mathrm{var}2}} + \ldots + coef(f_nc_k)\cdot\frac{valor\_\mathrm{var}n-\overline{X}_{\mathrm{var}n}}{s_{\mathrm{var}n}}$$

Sin embargo, muchos de los diversos paquetes estadísticos que existen en la actualidad facilitan todos estos cálculos. Algunos de ellos son *Statgraphics*, *SAS*, *SPSS,* Minitab, entre muchos más. Estos programas permiten la ejecución de una gran cantidad de técnicas estadísticas como lo es el análisis de componentes principales, el cual ha sido utilizado con más frecuencia desde la aparición de los ordenadores pues facilitan los numerosos cálculos que él implica. No obstante, este último paso se hace necesario solamente cuando las puntuaciones factoriales se utilizan como entrada en la aplicación de otra metodología o con otro fin. En algunos casos como en el de este estudio, las puntuaciones factoriales no fueron calculadas ya que el ACP se utilizó sólo para definir las variables que se usaron para la clasificación con MVS. Es decir, luego de haber definido las clases arrojadas por el ACP, se procedió a determinar a cuál de cada una de esas clases pertenecía cada una de las familias pero a partir de los datos originales.

Capítulo 5

Máquinas de Vectores Soporte

Este capítulo pretende destacar algunas de las características más importantes de las Máquinas de Vectores Soporte, explicando tanto fundamentos teóricos como matemáticos esenciales para el entendimiento de dicha técnica. Cabe destacar que las MVS fueron las que permitieron identificar los patrones de consumo de los venezolanos, al ser utilizadas en el estudio como técnica de clasificación no lineal.

5.1 Introducción a las MVS

Las Máquinas de Vectores Soporte (MVS), también conocidas como máquinas de soporte vectorial, están referidas a una técnica de aprendizaje automático que ha estado ganando popularidad en los últimos años. Los fundamentos teóricos de las MVS empezaron a ser desarrollados por Vapnik y otros autores desde la década de los 70. Sin embargo, no es hasta la década de los 90 cuando Vapnik, en compañía de Boser, Guyon y otros, presentan el modelo formal de las MVS como se conoce hoy día y se pudo comenzar a aplicar en problemas reales de reconocimiento de patrones. En la actualidad, las máquinas de vectores soporte pueden ser utilizadas para resolver problemas tanto de clasificación como de regresión. Algunas de las aplicaciones de clasificación o reconocimiento de patrones son: reconocimiento de firmas, reconocimiento de imágenes como rostros y categorización de textos. Por otro lado, las aplicaciones de regresión incluyen predicción de series de tiempo y problemas de inversión en general.

Las MVS han sido desarrolladas como una técnica robusta para clasificación y regresión para grandes conjuntos de datos complejos con ruido, es decir con variables inherentes al modelo, que para otras técnicas aumentan la posibilidad de error en los resultados pues resultan difíciles de cuantificar y observar. La teoría de generalización y las

funciones *kernel* son dos características claves de las MVS que le dan a gran utilidad ya que permiten identificar correctamente observaciones no conocidas y manejar datos no lineales sin necesidad de que se requiera conocer algún algoritmo no lineal explícito.

Las MVS "pertenecen a la familia de clasificadores lineales puesto que inducen separadores lineales o hiperplanos en espacios de características de muy alta dimensionalidad (introducidas por funciones núcleo o *kernel*) con un sesgo inductivo muy particular (maximización del margen)", (Carreras, Márquez y Romero, 2004). Sin embargo, la formulación matemática de las Máquinas de Vectores Soporte varía dependiendo de la naturaleza de los datos; es decir, existe una formulación para los casos lineales y, por otro lado, una formulación para casos no lineales. Es importante tener claro que, de manera general para clasificación, las máquinas de vectores soporte buscan encontrar un hiperplano óptimo que separe las clases.

5.2 Ventajas de las MVS

Las Máquinas de Vectores Soporten tienen ciertas características que las han puesto en ventaja respecto a otras técnicas populares de clasificación y/o regresión. Una de dichas características que vale la pena mencionar es que las mismas pertenecen a las disciplinas de aprendizaje automático o aprendizaje estadístico. La idea que hay detrás de este tipo de aprendizaje es la de hacer que las máquinas puedan ir aprendiendo, a través de ejemplos, las salidas correctas para ciertas entradas. Cuando los ejemplos son pares de entrada/salida el aprendizaje es conocido como aprendizaje supervisado. Dichos pares generalmente reflejan una relación funcional que mapea las entradas a las salidas. Los ejemplos que se utilizan para "enseñar" a la máquina se conocen como conjunto de entrenamiento. La metodología de aprendizaje permite que las MVS puedan resolver problemas sin necesidad de conocer la distribución o naturaleza de los datos con los que se desea trabajar. Además, esta característica permite obviar gran parte del trabajo del diseño y programación inherente a ciertas metodologías tradicionales.

Por otra parte, algunos autores como Gunn (1998) afirman que las MVS están ganando popularidad debido a sus tantas características atractivas, entre las cuales destaca la incorporación del principio de Minimización de Riesgo Estructural, el cual se 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 limite 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.

5.3 Definiciones básicas para las MVS

Existen diversos conceptos previos que se deben tener claros antes de entender los fundamentos matemáticos de las MVS. Algunos de dichos conceptos se presentan a continuación:

5.3.1 Dimensión VC

La dimensión VC (Vapnik-Chervonenkis) es un concepto fundamental de la teoría de aprendizaje estadístico, la cual es definida como la cardinalidad del mayor conjunto de puntos que cierto algoritmo de clasificación puede separar. En otras palabras, la dimensión VC es un escalar que representa la capacidad de ciertas funciones para separar un conjunto de datos (puntos) de entrenamiento. De esta manera, si la dimensión VC es h, existe por lo menos un conjunto de h puntos que pueden ser separados.

5.3.2 Maximización del margen

La maximización del margen es la idea que corresponde con exactitud a la aplicación del principio de Minimización de Riesgo Estructural. La maximización del margen, definido como la distancia de las muestras de entrenamiento a la frontera de decisión, se refiere a la selección del hiperplano separador que está a la misma distancia de los ejemplos más cercanos de cada clase. Carreras, Márquez y Romero (2004) señalan que lo anterior es equivalente a decir que se debe encontrar el hiperplano separador que está a la misma distancia de los ejemplos más cercanos de cada clase, donde los ejemplos se refieren a los datos utilizados para el entrenamiento. Además, mencionan que dicho hiperplano sólo considera los vectores soporte, es decir, aquellos puntos que están en las fronteras de la región de decisión, que es la zona donde puede haber dudas sobre a qué clase pertenece un ejemplo.

El modelo más simple de las MVS, conocido como clasificador de margen máximo, funciona sólo para datos linealmente separables en el espacio de características, por lo que no

puede ser usado en muchas aplicaciones de la vida real. Sin embargo, tal como lo afirman Cristianini y Shawe-Taylor (2000), es el algoritmo más fácil de aprender y forma la base fundamental para MVS más complejas. En la figura 5.1 se muestra geométricamente el hiperplano de margen máximo. La maximización del margen se encuentra dentro de la teoría de aprendizaje estadístico, específicamente dentro del principio de minimización de riesgo estructural.

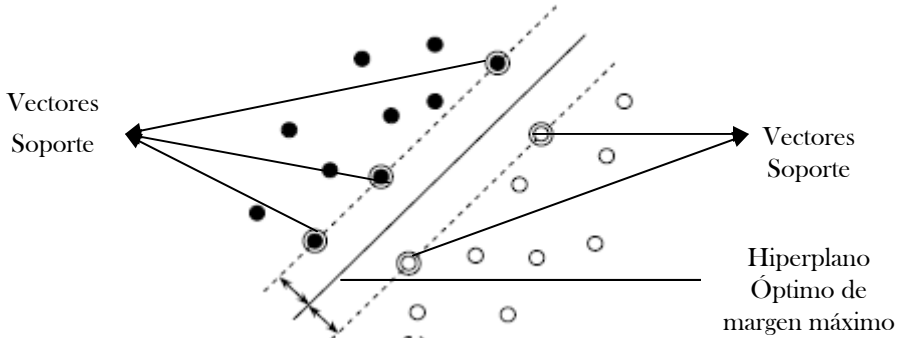


Figura 5.1 Hiperplano de margen máximo

5.4 MVS para el caso lineal

El modelo de MVS lineal con margen máximo es el caso más sencillo de clasificación de dicha técnica. Este caso se basa en el supuesto de que el conjunto de datos es linealmente separable en el espacio de entrada, es decir que, sin hacer ninguna transformación de los datos, los ejemplos pueden ser separados por un hiperplano de manera que en cada lado del mismo sólo queden ejemplos de una clase. En términos matemáticos, esto es equivalente a decir que existe un hiperplano $h: X \to \Re$ tal que $h(x) > 0$ para los ejemplos de la clase +1 y $h(x) < 0$ para los ejemplos de la clase -1. Un ejemplo de este caso se puede observar en la figura 5.2.

La distancia de un vector x a un hiperplano h, definido por (ω, b) como $h(x) = <\omega, b> + b$ viene dada por la fórmula $dist(h, x) = |h(x)| / \|\omega\|$, donde $\|\omega\|$ es la norma en $\Re^D$ asociada al producto escalar. Así pues, el hiperplano equidistante a dos clases es el que maximiza el valor mínimo de $dist(h, x)$ en el conjunto de datos. Como el conjunto es linealmente separable, se puede reescalar ω y b de manera que la distancia de los vectores más cercanos al hiperplano sea $1/\|\omega\|$. De esta manera, el problema de encontrar el

hiperplano equidistante a dos clases se reduce a encontrar la solución al siguiente problema de optimización:

$$\begin{aligned} &\text{Maximizar } 1/\|\omega\| \\ &\text{sujeto a: } y_i(<\omega, x_i> + b) \geq 1 \qquad (5.1) \\ &1 \leq i \leq N \end{aligned}$$

Escrito de otra manera, el problema puede ser resuelto a través de la siguiente formulación:

$$\begin{aligned} &\text{Minimizar } \frac{1}{2} <\omega, \omega> \\ &\text{sujeto a: } y_i(<\omega, x_i> + b) \geq 1 \qquad (5.2) \\ &1 \leq i \leq N \end{aligned}$$

Es importante destacar que cualquiera que sea el caso que se presente para la aplicación de las MVS, el problema a resolver es de optimización con restricciones y puede ser resuelto como un problema de programación cuadrática. Particularmente, el hiperplano separador de margen máximo ha demostrado una muy buena capacidad de generalización en numerosos problemas reales (Carreras, Márquez y Romero, 2004).

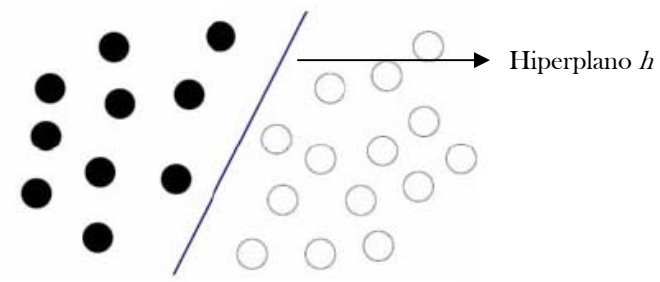


Figura 5.2 Clasificación de MVS para el caso lineal

5.5 MVS para el caso no lineal

Como se mencionó anteriormente, la mayoría de los casos de la vida real son de carácter no lineal y, por lo general, incorporan una gran cantidad de información necesaria para ser resuelta. Una de las grandes ventajas de las MVS es que éstas pueden ser utilizadas para resolver problemas no lineales mediante la utilización de funciones núcleo o *kernels*, las cuales permiten la solución sin necesidad de utilizar explícitamente algoritmos no lineales. Sin embargo, existen dos casos principales para la solución de datos que no son linealmente separables en el espacio de entradas; el primero, cuando los datos pueden ser separables con margen máximo pero en un espacio de características y, el segundo, cuando no es posible separar dichos datos linealmente incluso en un espacio de características. Ambos casos se explicarán en las subsecciones siguientes.

5.5.1 MVS con margen máximo en el espacio de características

Existen casos en los cuales los datos, debido a su naturaleza, no pueden ser separados linealmente a través de un hiperplano óptimo, que es la idea que hay detrás de margen máximo. Sin embargo, en muchas situaciones los datos, a través de una transformación no lineal del espacio de entradas, pueden ser separados linealmente pero en un espacio de características, en el cual se pueden aplicar los mismos razonamientos que para las MVS lineal con margen máximo.

Aunque la dimensión del espacio de características puede ser bastante grande, para ciertas transformaciones y ciertos espacios de características, se puede calcular el producto escalar usando las denominadas funciones núcleo. En la figura 5.3 se puede observar un ejemplo donde se puede encontrar un hiperplano óptimo en el espacio de características, a través de un *kernel.*

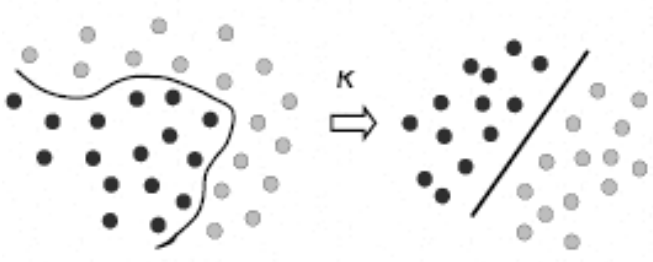


Figura 5.3: Clasificación de MVS para el caso no lineal

Una función núcleo o *kernel* se puede definir como aquella que permite realizar la separación y el traslado de los datos al espacio de características. Existen diversos *kernels* predeterminados conocidos entre los cuales se destaca el lineal, el RBF (Función de Base Radial), el polinomial, el sigmoidal, entre otros más. Matemáticamente, un *kernel* se puede definir como una función $K : X \times X \to \Re$ tal que $K(x, y) = < \Phi(x), \Phi(y) >$, donde Φ es una transformación de X en un cierto espacio de Hilbert $\Im$. Carreras, Márquez y Romero (2004) afirman que la función núcleo permite calcular el producto escalar $< \Phi(x), \Phi(y) >$ en el espacio de características sin necesidad de usar ni conocer la transformación Φ.

Luego de conocer la idea de lo qué es una función núcleo y qué se puede lograr con ella, se puede proceder a especificar el problema de optimización a resolver para las MVS con margen máximo en el espacio de características. El problema de programación cuadrática con restricciones a resolver es el siguiente:

$$\text{Maximizar } \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \alpha_i \alpha_j K(x_i, x_j)$$

$$\text{Sujeto a: } \sum_{i=1}^{N} y_i \alpha_i = 0 \qquad (5.3)$$

$$\alpha_i \geq 0, \qquad 1 \leq i \leq 0$$

Donde $K(x, y)$ es la función núcleo.

5.5.2 MVS con margen blando

Las MVS con margen blando se refieren a aquellos casos en los cuales los datos no son linealmente separables, ni siquiera en el espacio de características. También es necesario este enfoque cuando los datos pueden ser clasificados linealmente en el espacio de características pero las soluciones son sobreajustadas a los ejemplos y, con esto, se tiende a tener mala generalización. Este modelo es el que precisamente funciona bien con la inclusión de ruido pues resulta ser mucho más robusto que los anteriores modelos explicados.

Las MVS con margen blando incorporan variables de holgura al modelo permitiéndole a las máquinas seleccionar un clasificador con cierto margen de error pero que a su vez tienda a tener una mejor generalización. En otras palabras, estas variables, cuyo valor será siempre positivo o igual a cero, permitirán que con cierto margen no se cumplan las restricciones

estrictamente. Además, será necesario incluir tantas variables de holgura como datos se tengan, pues para cada instancia de datos se tendrá su respectiva variable de holgura.

La inclusión de variables de holgura a las restricciones en el modelo debe equilibrarse incluyendo en la función objetivo un término de regularización que depende de dichas variables. Este término es el que se conoce como el parámetro C, el cual determina la holgura del margen blando; es decir, el parámetro C es el que permite al clasificador un cierto margen error en el momento de clasificación. Con esto, el problema a optimizar resulta muy similar al de los modelos con margen máximo, con la diferencia de que en este modelo de margen blando se incluye un término en la función objetivo dependiente del parámetro C y las variables de holgura y, a su vez, a cada restricción se le añadirá las respectivas variables de holgura. Además, se incluirá una nueva restricción que es la que limita el valor del parámetro C.

5.6 Fases para multiclasificación con MVS

Los problemas de multiclasificación pueden ser resueltos a través de una serie de pasos utilizando *softwares* especiales para esto, entre los cuales se encuentran: SVM^{light}, LIBSVM, diferentes *toolbox* de Matlab, Spider, WEKA, entre muchos más. La utilización de los programas para llevar a cabo multiclasificación y biclasificación facilitan el uso de MVS, ya que sin ellos sería bastante complejo e incluso imposible la aplicación de esta técnica. A continuación se explican, de manera general, pasos fundamentales para efectuar clasificación utilizando MVS a través de un programa diseñado para ese fin. Cabe destacar que éstos pasos pueden variar, entre otras cosas, del *software* que se esté utilizando y de la complejidad de los datos con los que se esté trabajando.

5.6.1 Selección de las muestras

El primer paso fundamental para llevar a cabo clasificaciones con MVS, al igual que con otras técnicas, se refiere a la selección de la muestras. Para esto es necesario determinar cuáles observaciones o conjunto de datos se utilizarán para el entrenamiento y cuáles para las pruebas. Dicha selección se puede hacer de manera aleatoria siempre y cuando se garantice que dentro de cada conjunto se esté incluyendo diversidad de los datos con los que se esté trabajando. Por ejemplo, en el caso de la presente investigación, donde se estudiaron patrones de consumos de venezolanos, es importante que, tanto para el entrenamiento como para las

pruebas, se incluyan observaciones de todas las regiones consideradas. Por lo tanto, fue necesario recurrir al muestreo aleatoria estratificado.

Por otro lado, algunos de los *softwares* disponibles para clasificación con Máquinas de Vectores Soporte brindan herramientas de selección de muestras y pruebas de acuerdo a diferentes criterios. Esto se podrá ver con más detalle en la fase donde se selecciona el tipo de prueba para el clasificador, con el cual se podrán seleccionar las muestras de prueba de diferentes maneras.

5.6.2 Transformación de los datos al formato del *software*

El primer paso esencial, luego de haber seleccionado el *software* con el cual se procederá a realizar la clasificación, consiste en transformar los datos al formato compatible con dicho *software* seleccionado. El formato que deben tener los datos depende exclusivamente del programa que se utilizará; algunos usan, por ejemplo, tablas de Excel como entrada para el entrenamiento y prueba de las MVS. Sin embargo, es necesario tener en cuenta que las MVS requieren que cada instancia de los datos esté representada como un vector de números reales. Por lo tanto, si existen atributos categóricos o cualitativos, éstos deben ser previamente transformados a datos numéricos.

5.6.3 Selección del *kernel*

La selección del modelo consiste en determinar el *kernel* que se utilizará, los valores de los parámetros del mismo y el tipo modo de evaluación del clasificador. En este momento se hace oportuno tener claro cuáles son los *kernels* que se pueden utilizar, los cuales varían dependiendo del *software* que se haya decidido manipular. Además, es necesario conocer cuáles son los parámetros que pueden variar en cada *kernel,* los cuales dependen directamente del tipo de *kernel,* pues cada uno posee diferentes parámetros asociados a él.

Aunque existen varios tipos de *kernels* comunes, es necesario decidir cuál de ellos se probará primero. Esta es una etapa fundamental 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. Muchos autores como Hsu *et al.* (2007), Moro y Hurtado (2006), recomiendan el uso del *kernel* de Función de Base Radial (RBF) como primera opción, pues ha demostrado en muchas ocasiones un buen desempeño y, con esto, se han obtenido resultados bastante satisfactorios. Una de las ventajas que tiene el RBF es

que es un *kernel* no lineal, que a diferencia del *kernel* lineal, mapea de manera no lineal los ejemplos a un espacio de características de mayor dimensión y, con esto, maneja casos en los que la relación entre clases y atributos son no lineales. Además, el RBF posee menos parámetros que otros *kernels,* como el polinomial, lo que reduce la complejidad de la selección del modelo. Sin embargo, es recomendable realizar varias corridas de entrenamiento y prueba con diferentes *kernels* para observar cómo varía el desempeño de la técnica y poder determinar cuál resulta ser el mejor *kernel* para los datos estudiados.

5.6.4 Selección del tipo de prueba para el clasificador

Como se mencionó anteriormente, la elección del *kernel* además de incluir la selección del tipo, incluye la determinación de los valores de los parámetros asociados. Por lo tanto, las corridas que se realizan deben alternar tanto el tipo de *kernel* como los valores de los parámetros de cada uno de ellos. Por otra parte, los resultados de las diferentes corridas dependen de la manera cómo se pruebe el clasificador. Las opciones para probar el desempeño del clasificador son: 1) Usar el conjunto de entrenamiento; 2) Suministrar un conjunto de datos para la prueba distinto del de entrenamiento; 3) Evaluación con validación cruzada; y 4) Usar un porcentaje de los datos suministrado para el entrenamiento y el resto para la prueba. El uso de conjunto de entrenamiento para realizar la prueba no resulta lógico cuando se desea medir la capacidad predictiva de las MVS pues se probará con las mismas instancias conocidas con las que fueron entrenadas.

La evaluación a través de validación cruzada resulta bastante buena y es altamente recomendada por diversos autores como *Hsu et al.* (2007). Ésta consiste en suministrarle al *software* todos los datos estudiados, el cual los dividirá en un número especificado y realizará tantas evaluaciones como ese número tomando una parte como el conjunto de evaluación y el resto como datos de entrenamiento. Por ejemplo, si se decide validación cruzada con parámetro 10, el *software* dividirá el conjunto de datos en 10 subconjuntos de similar tamaño y, luego, tomará 9 de los subconjuntos para entrenar las MVS y el subconjunto restante se usará para realizar la prueba del clasificador; de esta manera, variará la combinación y realizará 10 entrenamientos con sus respectivas 10 pruebas.

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.

5.6.5 Entrenamiento

Una vez que se ha seleccionado el tipo de *kernel* óptimo de acuerdo a la naturaleza de los datos estudiados, los valores óptimos de sus parámetros asociados y el tipo de prueba que se usará, se puede proceder a hacer el entrenamiento formal de las MVS. La tarea de clasificación involucra tanto entrenamiento como pruebas de instancias de datos. Cada instancia en el conjunto de entrenamiento debe contener un valor objetivo o etiqueta de clase, es decir la clase a la cual pertenece dicha instancia. Así, lo que se hace en esta es especificar cuál es el conjunto de datos que se usará para el entrenamiento, es decir, los ejemplos que se le darán a las MVS para que ellas "aprendan" a clasificar correctamente las instancias con el menor error de clasificación posible.

5.6.6 Pruebas

La última fase de las MVS consiste en realizar las respectivas pruebas al clasificador seleccionado para medir su desempeño. Como se mencionó anteriormente, existen diferentes formas de realizar las pruebas, ya sea a través de validación cruzada o, simplemente, utilizando un porcentaje de los datos proporcionados. 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 producir un modelo que permita predecir correctamente la clase a la que pertenece cada instancia.

Esta fase es de gran importancia ya que a partir de los resultados que se obtengan, se puede inferir sobre los mismos respecto a diferentes medidas de evaluación de desempeño de las MVS. Uno de los factores que se observa en esta etapa es la habilidad de los modelos probados para clasificar correctamente instancias desconocidas. En otras palabras, lo que se hace es evaluar la efectividad de los parámetros seleccionados para producir clasificaciones acertadas. Además, en esta fase es donde se estudia la consistencia de dichos modelos.

5.6.7 Interpretación de los resultados

Una buena interpretación de los resultados obtenidos a través de la aplicación de la técnica se hace fundamental pues sin ella el estudio pierde sentido. Siempre es necesario interpretar los resultados que se obtienen con la idea de darle significado a los "valores" obtenidos, para lo cual se requiere que el investigador conozca a fondo los datos que se están estudiando. El análisis de los resultados de las MVS incluye la identificación del error de clasificación que se tiene con el modelo seleccionado, las clases establecidas por dicho modelo, etc. En muchos casos, la interpretación incluirá comparaciones con otras técnicas estudiadas con la intención de verificar la capacidad predictiva de las MVS.

Capítulo 6

Preprocesamiento de los datos

El preprocesamiento de los datos puede ser visto como una etapa en la que los datos de una cierta muestra son manipulados y transformados a un conjunto de datos más sencillo y conciso, el cual pueda ser interpretado y utilizado de manera más eficiente para un posterior análisis. Cabe destacar que este paso preparatorio resulta fundamental para el investigador ya que le permite familiarizarse con el conjunto de datos y, de esta manera, hacer mejores análisis e interpretaciones futuras que se requieran.

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.

Además, la preparación de los datos puede generar un conjunto más pequeño que el original, lo cual puede mejorar la eficiencia de un posterior análisis a través de la aplicación de alguna técnica. Esto se logra a través de la selección relevante de los datos (eliminación de registros repetidos, anomalías, etc.) y la reducción de los mismos (selección de características y variables a utilizar, eliminación de información irrelevante para el estudio, etc.). Igualmente, la preparación de los datos permite mejorar la calidad de los datos mediante la recuperación de datos faltantes, eliminación de valores atípicos, entre otros. 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.

Existen diversos pasos que se deben seguir para hacer un buen preprocesamiento de datos. Sin embargo, estos pasos a seguir varían de acuerdo a diferentes autores. Para llevar a cabo el preprocesamiento de este estudio se efectuaron los pasos básicos y comunes para la mayoría de los autores estudiados. No obstante, antes de realizar cualquier modificación a los datos es necesario ligarse muy bien con ellos; es decir, conocer la manera cómo se encuentran estructurados, las variables incluidas y la manera cómo se definen las mismas, el tamaño de la muestra, su distribución, etc. En las secciones siguientes se explicará detalladamente el preprocesamiento realizado paso a paso.

6.1 Estructura de la muestra de los datos originales

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; cada una de ellas con distinta información respecto a los resultados de la II Encuesta Nacional de Presupuestos Familiares. Para conocer y entender el contenido de las tablas fue necesario recurrir al catálogo de las variables que se contemplan en la encuesta, el cual fue realizado por el Banco Central de Venezuela y se conoce como Diseño de Registro. Las 23 tablas contenidas, mencionadas anteriormente, son las siguientes: 1) Actividad; 2) Categorías; 3) Continua; 4) Descripción de consultas; 5) Dotación; 6) Entidades; 7) Gastos diarios del hogar; 8) Gastos diarios personales; 9) Gastos del hogar en restaurantes; 10) Gastos mensuales; 11) Gastos personales en restaurantes; 12) Gastos trimestrales y anuales; 13) Ingresos; 14) Miembros; 15)Ocupación; 16)Región; 17)Transferencias; 18) Vehículo; 19) Vivienda02; 20)Vivienda03; 21)Vivienda04; 22)Vivienda06; y, por último 23)Vivienda.

Como se mencionó anteriormente, en cada una de las tablas 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. Por ejemplo, la tabla 3 denominada en la base de datos como "continua" contenía información respecto al número de hogares que habitaban en la misma vivienda e indica si cada una de las viviendas estaba dotada, entre otras cosas, con cocinas eléctricas o a gas (directo o con bombonas), antena parabólica, televisión

por cable y teléfonos. Por otro lado, las tablas vivienda, vivienda02, vivenda03, vivienda04 y vivienda06 contenían información diversa respecto a características de las viviendas. Por ejemplo, en la tabla vivienda02 se especificaba el número de miembros que habitaban en cada vivienda y los nombres de los mismos. El tipo de vivienda, de piso, de techo y de paredes; así como el número de habitaciones y baños; la afirmación o negación de si las viviendas gozaban de servicio de agua, alumbrado público y vigilancia privada; son datos encontrados en la tabla "vivienda04".

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. Cabe destacar que las variables que se escogieron, respecto a las familias venezolanas, fueron: los desembolsos en cada uno de los grupos de gastos que se establecerían y el ingreso total. Además, se decidió incluir información respecto a la entidad a la que pertenece cada vivienda y el número de miembros que habitan en las mismas.

6.2 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. A continuación se detallará cuáles fueron las modificaciones realizadas, justificándolas, y la manera cómo se compusieron los grupos de gastos finales con los que posteriormente se harían diferentes análisis. Los códigos y detalles de la composición de dichos grupos se muestran en el anexo I. Los nuevos 22 grupos de gastos quedaron establecidos de la siguiente manera:

1) Alimentos y bebidas no alcohólicas: en los datos obtenidos para la realización del estudio existe un grupo denotado "Alimentos, bebidas y tabaco". Se decidió descomponer este grupo en grupos más pequeños en los cuales se incluían los alimentos y las bebidas no alcohólicas en un grupo, las bebidas alcohólicas y tabaco en otro grupo y, por otro lado, los alimentos y bebidas no alcohólicas tomadas fuera del hogar. Esta decisión está fundamentada en la importancia de estudiar estos gastos por separado ya que, para efectos del estudio de patrones de consumo, resulta esencial diferenciar entre bienes básicos (alimentos y bebidas no alcohólicas) y bienes innecesarios (bebidas alcohólicas y tabaco). Además, resulta interesante diferenciar el consumo de las familias venezolanas en restaurantes. De esta manera, el grupo 1 quedó formado por todos los subgrupos de la clasificación original que estaban referidos a cualquier tipo de alimento (Cereales y sus productos derivados, Carnes de ganado vacuno, Carne de aves, Frutas, Hortalizas, Leche y sus derivados, Alimentos especiales para niños, etc.). Las bebidas no alcohólicas (agua mineral, batidos de frutas, gaseosas, etc.) también se incluyeron en este mismo grupo debido a que éstos comúnmente son utilizados para acompañar los alimentos y, además, también pueden ser consideradas como bienes básicos o necesarios.

2) Bebidas alcohólicas y tabaco: como se mencionó anteriormente, este grupo se desprende del grupo inicial denominado "Alimentos, bebidas y tabaco". El Grupo 2 quedó compuesto por los siguientes subgrupos originales: Bebidas alcohólicas (aguardiente, cerveza, ron, etc.); Bebidas alcohólicas tomadas fuera del hogar; y, por último, el subgrupo Tabaco (cigarrillos, chimó, picadura, etc.).

3) Comidas fuera del hogar: Para efectos del presente estudió se consideró conveniente separar los alimentos y bebidas no alcohólicas tomados fuera del hogar de aquellos que se compran para preparar y comer en el hogar. Esto se debe a que resulta de interés observar el comportamiento de consumo del venezolano respecto a comidas en restaurantes y establecimientos similares. Es fundamental destacar que no se incluyó este rubro dentro del grupo Recreación y Esparcimiento ya que no necesariamente los venezolanos decidimos comer en algún establecimiento de venta de comida por algunas de estas razones. A menudo ocurre que las personas comen fuera del hogar por la lejanía del hogar al trabajo, falta de tiempo, etc.

4) Vestido y calzado: este grupo es equivalente al mismo grupo Vestido y Calzado de los datos originales obtenidos para realizar el presente estudio. Dicho rubro se mantuvo bajo la misma clasificación ya que para estudios de consumo resulta bastante representativo. En este grupo están incluidos todos los gastos en vestimenta, calzado y accesorios para cualquier miembro del hogar. Asimismo, están incluidos gastos en servicios y materiales para la confección, reparación y prendas de vestir y calzado.

5) Vivienda y sus servicios: este grupo está definido para el estudio de la misma forma que en los datos originales, bajo el mismo nombre, con la diferencia de que se excluyó el subgrupo definido como "Mantenimiento de la vivienda", el cual incluye gastos de remodelación, materiales para la construcción, entre otros. El nuevo grupo Vivienda y sus servicios incluye gastos de alquiler, gastos en servicios básicos como electricidad, gas, agua, etc. Igualmente, se consideró conveniente incluir gastos en servicios domésticos y gastos en seguros relacionados con el hogar debido, fundamentalmente, a la relación que tienen con los demás gastos incluidos. Estos gastos mencionados pertenecían, originalmente, a los grupos "Mobiliario, Equipos del Hogar y Mantenimiento Rutinario de la Vivienda" y "Otros gastos", respectivamente.

6) Mobiliario y equipos del hogar: esta categoría de gastos abarcó todos los desembolsos en los que incurren miembros de familias que están relacionados con muebles, electrodomésticos, utensilios y equipos para el hogar. Resultó de desglosar un mayor grupo llamado "Mobiliario, equipos del hogar y mantenimiento rutinario de la vivienda". La exclusión de la última parte (mantenimiento rutinario de la vivienda) se consideró necesaria debido al tipo de gastos a los que se referían. Los muebles y los equipos del hogar son gastos que, generalmente, se hacen con menos frecuencia y su fin consiste en equipar la vivienda, entre otras cosas, para mayor comodidad. Mientras que los gastos rutinarios de la vivienda se hacen con mucha más frecuencia y son necesarios para mantener la vivienda en el mejor estado posible.

7) Mantenimiento de la vivienda: como se señaló anteriormente, es el equivalente a una parte del grupo original "Mobiliario, equipos del hogar y mantenimiento rutinario de la vivienda". En este conjunto se contemplaron todos aquellos gastos que se realizan para sostener en "buen" estado las viviendas, tanto los más básicos y comunes (artículos de limpieza, bolsas de basura, bombillos, servilletas, servicios de fumigación, servicios de reparación de alfombra y otros bienes del hogar, etc.) como los más eventuales (remodelaciones y mantenimiento mayor).

8) Salud: este grupo es el mismo precisado en los datos que fueron obtenidos para la realización del estudio. En él se incluyen todos los gastos en salud afines con medicinas en general, artículos médicos como aparatos ortopédicos, sillas de ruedas, enfermeras, clínicas, exámenes de laboratorio, cirugías, etc. Además, están incluidos aquellos gastos en seguros y previsión social relacionados con la salud.

9) Transporte personal: originalmente formaba parte de un grupo denominado "Transporte y comunicación". Fue extraído y considerado un rubro separado de otros que, de igual manera, formaban parte de dicho grupo original. El desglose del grupo Transporte y comunicación se debió a la importancia de reflejar los consumos en los diferentes tipos de transporte y

las distintas maneras de comunicación existentes. El transporte personal contempla todos aquellos gastos en medios de transporte que pueden ser adquiridos para uso personal como vehículos y, además, los gastos necesarios para su mantenimiento tanto en servicios como en piezas y accesorios. Al mismo tiempo, se incluyeron gastos en seguros como el de responsabilidad civil que están relacionados directamente con el transporte personal.

10) Transporte colectivo: considera todos los gastos en cualquier tipo de transporte colectivo tanto privado como público. Algunos de los más comunes son mensualidades de transporte escolar, pasajes en cualquier tipo de transporte público (autobuses, carros por puesto, rústicos, metro, etc.) y carreras en taxis.

11) Telefonía fija: como se señaló anteriormente, resulta importante reflejar en este estudio el consumo en los diferentes medios existentes para la comunicación. La telefonía fija, resulta entonces un rubro importante de destacar en el que se incluyen gastos tanto en los módems como en los servicios de llamadas y reparación de teléfonos.

12) Telefonía móvil: la telefonía fija implica un gasto muy importante para el estudio en patrones de consumo ya que, como puede observarse en la vida cotidiana, es un servicio al cual acceden gran parte de los venezolanos. En los gastos en telefonía móvil están contenidos aquellos en celulares, mensualidades por telefonía celular, los servicios de reparación, entre otros. La idea que hay detrás de esta separación es observar gastos en telefonía móvil e Internet por separado debido a la importancia que se le ha estado dando en los últimos años. A pesar de que para la fecha no existía esta tendencia notoria en los venezolanos y en la población mundial en general, resulta interesante la separación para observar el consumo en estas tecnologías para aquél entonces y, además, para comparaciones con posteriores estudios que se hagan en este ámbito.

13) Internet: el número de usuarios que acceden a Internet crece continuamente a medida que crece su importancia y utilidad. Hoy en día Internet resulta la

manera más cómoda, fácil y rápida de comunicarse a nivel mundial. Es por esto y mucho más que se decidió que este modo de comunicación fuese considerado como un rubro separado de los demás. En este grupo se contemplan todo tipo de gastos en servicios de Internet.

14) Otras comunicaciones: en este grupo están incluidos todos aquellos gastos que se realizan en otros modos de comunicación más clásicos como servicio de correspondencia, fax, telegramas, etc.

15) Educación y cultura: este rubro formaba parte de uno más extenso denominado "Educación, cultura y esparcimiento". Al igual que se hizo con otras secciones iniciales, dicho rubro fue desglosado en grupos más específicos. Particularmente, la necesidad de observar el consumo en educación y cultura separado del consumo en esparcimiento se fundamenta en la diferencia obvia existente en los diferentes gastos de acuerdo a la importancia de ellos. En otras palabras, es significativo diferenciar la educación y la cultura de las actividades que se realizan con fines de distracción, esparcimiento y/o diversión. Dentro del grupo Educación y cultura se hallan artículos necesarios para la educación como libros, diccionarios, enciclopedias, computadoras y, además, matrículas de inscripción por educación de distintos niveles, las mensualidades, etc.

16) Recreación y esparcimiento: incluye parte del rubro Educación, cultura y esparcimiento. Como se mencionó anteriormente, se consideró relevante separar los gastos en educación y cultura de los gastos en esparcimiento, recreación y/o diversión. Para esto fue necesario revisar detenidamente cada uno de los gastos que pertenecían al rubro Educación, cultura y esparcimiento y establecer a cuál nuevo grupo pertenecería. Se consideraron apropiados a estar en este grupo gastos en: artículos para manualidades, juguetes, juegos de diferentes tipos, artículos deportivos, radios, televisores, cámaras, entradas a espectáculos, entradas a museos y zoológicos, entre muchos más.

17) Artículos, efectos personales y sus servicios: es el mismo designado en los datos primarios con el mismo nombre. La razón por la cual se decidió

mantener tanto el grupo como su nombre es porque se consideró que, para los fines que perseguía el estudio, resultaba suficientemente representativo. Aquí están contenidos todos los gastos en cuidado e higiene personal como cepillos de dientes, cortes de cabello, cosméticos, productos para el cabello, perfumes y, además, los gastos relacionados con el aspecto personal como relojes, llaveros, lentes de sol, prendas, etc. Asimismo, se incluyen los gastos relacionados con materiales para la fabricación y reparación de artículos personales.

18) Gastos financieros: los productos y servicios contenidos en este rubro pertenecían a uno mayor conocido como "Otros gastos". Debido a la variabilidad de gastos incluidos en este último grupo inicial, éste fue descompuesto en tres grupos más específicos. El primero de ellos es el que se describe en esta parte y está conformado por todos los gastos en servicios bancarios, en cajas de seguridad y otros relacionados con el manejo de las finanzas.

19) Gastos tributarios y legales: éste equivale al segundo de los grupos descompuestos de "Otros gastos". En él están incluidos los gastos en diferentes tipos de impuestos, en notarías, gastos por multas, expediciones de partidas de nacimiento, actas de matrimonio, cédulas, etc. Al mismo tiempo, están comprendidos los gastos en intereses sobre préstamos, ya que éstos pueden ser considerados como gastos legales.

20) Otros gastos: está formado, principalmente, por todos aquellos gastos que pertenecían al rubro primario "Otros gastos" pero que no figuraron para ser incluidos en ninguno de los dos anteriores. Aquí se contemplan gastos en compra de animales, alimentos para animales, aportes a instituciones, árboles ornamentales, servicios funerarios, arreglos florales, entre otros. Igualmente, están gastos pertenecientes a otros grupos originales que, por alguna razón particular, no podían estar en ninguno de los otros grupos establecidos. Algunos de estos gastos son en servicios de mudanzas, armas de fuego y baterías. Como puede observarse, algunos de los gastos de este grupo no pudiesen pertenecer a ninguno de los otros grupos establecidos y

no serían suficientemente significativos para crear otro grupo aparte. Por otro lado, existen algunos que pudiesen pertenecer a más de un grupo; por ejemplo, las baterías, que pudiesen pertenecer al grupo Recreación y esparcimiento o al grupo educación y cultura, entre otros más. Esto, mencionado anteriormente, se debe fundamentalmente a que no se puede saber el fin con el cual cada persona realizó el gasto.

21) Satélite y cable para TV: se desprendió del grupo original con el nombre "Transporte y comunicaciones". La razón por la cual se separó de los demás es que es un tipo de comunicación distinta a las que pertenecen a los otros grupos. Los gastos en estos rubros no se incluyen dentro de Vivienda y sus servicios ya que no puede considerarse como un servicio básico para la vivienda como los demás que están incluidos en este grupo.

22) Viajes: conformado por gastos de diferentes grupos iniciales. Los gastos que se contemplan en este grupo son todos aquellos relacionados con viajes en general. Debido a que no se puede saber el fin o la razón de los viajes (diversión, trabajo, estudio, etc.) se dejaron por separado. Algunos de los gastos incluidos en este grupo son en: servicios de alquiler de vehículos de transporte (alquiler de vehículos rústicos, vehículos de paseo, lanchas, botes, etc.); pasajes al interior y exterior del país; pasajes marítimos; hoteles; pagos de peajes; expedición y renovación de pasaportes y visas; y, por último, pagos por seguros de asistencia al viajero.

6.3 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 26 columnas en total y estaría estructurada como se muestra en la figura 6.1. Sin embargo, para llegar a tenerla de esa manera fue necesario realizar varios procedimientos, los cuales se describirán a continuación.

6.3.1 Totalización de los gastos por grupos

Como se refirió previamente, en los datos conseguidos para la elaboración del estudio, los gastos se encontraban distribuidos en varias de las tablas de la base de datos. Particularmente, todos los valores de los desembolsos dados a conocer por las familias encuestadas se localizaban en las tablas "gastos diarios del hogar", "gastos diarios personales", "gastos del hogar en restaurantes", "gastos mensuales", "gastos personales en restaurantes" y en "gastos trimestrales y anuales". Es notorio que los gastos estaban separados de acuerdo los períodos de referencia establecidos por el Banco Central de Venezuela. En este punto, es importante recordar que aquellos gastos comunes, con menores precios y que pueden ser olvidados generalmente por esas razones, fueron considerados como gastos diarios. Mientras que aquellos gastos que puedan tender a ser fáciles de recordar porque se hacen con menos frecuencia y son, por lo general, gastos costosos, calificaron como gastos anuales.

NUMERBCV	COD_ENTIDAD	NUM_MIEMBROS	GASTO G1	GASTO G2	...	GASTO G22	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 6.1: Estructuración de matriz de datos

Para poder agrupar los gastos de acuerdo a los diferentes grupos fue necesario llevarlos todos a un mismo período de referencia. Esto se debe a que, para una mejor identificación de patrones de consumo más precisos y de mejor calidad, es ventajoso que no exista tanta variabilidad en los datos. Además, resultaría ilógico sumar, por ejemplo, gastos mensuales con gastos anuales. Cabe destacar que el período de referencia seleccionado fue el mensual ya que resulta intermedio entre todos.

En gastos diarios del hogar y en gastos diarios personales estaba contenida información referente a las adquisiciones que realiza el hogar en bienes y servicios tales como: alimentos, bebidas, artículos de limpieza e higiene, medicamentos, transporte, combustible y otros gastos rutinarios. La diferencia entre ambas tablas radica en que la segunda contenía información de cada uno de los miembros mayores de 12 años que realizan gastos fuera del hogar, mientras que la primera es el informante calificado quien revela los desembolsos en los que ha

incurrido para el hogar. No obstante, dichos gastos corresponden a aquellos registrados por los encuestados durante la semana de referencia. Por lo tanto, juntos constituyen el gasto total semanal en el que incurrió cada hogar durante la aplicación de la encuesta. De esta manera, todos los gastos incluidos dentro de estas tablas fueron multiplicados por cuatro para ser llevados a gastos mensuales.

La tabla gastos mensuales contiene información sobre las adquisiciones efectuadas por el hogar en bienes y servicios cuya frecuencia de consumo sea aproximadamente mensual. En general, están referidos a gastos en prendas de vestir, calzado y accesorios personales, servicios cuyo pago se realiza mensualmente, artículos y accesorios del hogar. Debido a que estos gastos ya estaban expresados de acuerdo al período de referencia deseado, no fue necesario realizar sobre ellos ninguna operación para su posterior totalización con los demás.

Los gastos del hogar en restaurantes y gastos personales en restaurantes tuvieron un tratamiento similar a los gastos diarios, debido a que todos se referían a los gastos totales para la semana de referencia, es decir cada uno, de acuerdo a su grupo, constituye el gasto semanal en el que incurrió el hogar para ese período de referencia. En las tablas mencionadas se encontraron todos los desembolsos, tanto del hogar como personales, relacionados con consumo en restaurantes y lugares similares. Para llevar dichos gastos al período de referencia mensual, se multiplicaron por 4, pues cada mes tiene aproximadamente cuatro semanas.

La tabla donde estaban incluidos los gastos trimestrales y anuales requirió de un poco más de tiempo y esfuerzo debido a que estaban mezcladas adquisiciones en dos períodos de referencia distintos, que pudieron ser recordados por los encuestados en referencia al trimestre o al año anterior. Por ello, fue necesario, antes de todo, identificar cuáles de los gastos eran trimestrales y cuáles anuales. Para esto se refirió al cuestionario "EPF" relacionado con estos gastos, específicamente a la sección "Recordatorio de Gastos Trimestrales y Anuales", en el cual se listaban tipos de gastos de acuerdo a su período de referencia. En dicho recordatorio se señalaban como gastos trimestrales todos aquellos afines con muebles, equipos y accesorios para bebés, equipos electrodomésticos del hogar, equipos y artículos relacionados con la práctica del deporte, equipos de computación y de comunicación, equipos médicos y servicios relacionados con la salud, viajes de más de tres días de duración, entre otros. Los gastos anuales resultaron ser todos aquellos con valor elevado que facilita al encuestado recordarlos, entre los cuales estaban: gastos por remodelaciones de la vivienda, compra de vehículos, pago

de impuestos y seguros, servicios financieros e intereses pagados. Los gastos trimestrales y anuales fueron divididos entre 3 y 12, respectivamente, para ser llevados al período de referencia mensual. Y es que resulta lógico pensar que una tercera parte de un gasto trimestral podría ser atribuible a un mes y, análogamente, una doceava parte de un gasto anual se le podría atribuir a uno de 12 meses.

Después de haber hecho todas las operaciones necesarias para tener todos los gastos expresados de acuerdo a un mismo período de referencia (mensual), se pudo proceder a realizar la totalización correspondiente a cada uno de los grupos de gastos previamente establecidos. Con esto, se tendría la información relativa a 22 de las 26 columnas de la matriz de datos con la que, posteriormente, se aplicarían diferentes técnicas de análisis.

6.3.2 Totalización de los ingresos

Al igual que se hizo con todos los gastos, los ingresos también debían estar expresados en el período de referencia mensual y debían estar totalizados para cada uno de los hogares encuestados. La tabla "ingresos" proporcionó información que permitió medir los ingresos tanto mensuales como anuales que percibieron los miembros del hogar, por cualquier concepto, en determinados períodos de referencia. Fue la tabla de los códigos de los ingresos la que permitió distinguir cuáles de los ingresos eran reconocidos como ingresos mensuales y cuáles de ellos como ingresos anuales. Luego de conocidos los códigos que identificaban a los ingresos mensuales y los que identificaban a los anuales se procedió a transformar los últimos en ingresos mensuales. De esta manera, se obtuvieron todos los ingresos expresados en el período de referencia deseado y se pudo totalizarlos para todos los hogares. Cabe destacar que, inicialmente, se incluyeron los ingresos bajo el criterio de que éstos podrían ser utilizados para construir la variable de salida que podría definir las clases para la clasificación con MVS.

6.3.3 Unificación de información para la creación de la matriz de datos

Una vez que se obtuvieron los gastos (de cada hogar para cada grupo) y los ingresos mensuales totalizados para cada uno de los hogares, se creó la matriz tentativa de los datos. La unificación de las 22 columnas referidas a los 22 grupos de gastos establecidos mas la columna de ingresos, junto con los códigos que identifican a los hogares, la entidad a la que pertenece cada uno y el número de miembros que los conforman, permitieron la creación de la matriz de datos formada por 26 columnas en total. Sin embargo, esto no llegó hasta ahí ya que a partir de

ese momento fue necesario profundizar en el contenido de dicha matriz para evaluar los datos y poder determinar si existían valores atípicos y/o valores faltantes.

6.3.4 Manejo de valores faltantes

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; es decir, se eliminaron aquellas observaciones que no tenían esta información completa. 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.

6.3.5 Manejo de valores atípicos

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). A pesar de que los valores atípicos observados estaban muy por encima del resto de los valores en los mismos grupos de gastos, e incluso de todos los grupos, quizás sí hayan sido registrados por los encuestados. Sin embargo, dichos valores también pueden surgir de ciertos errores, por ejemplo, errores de trascripción. 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 no eliminar 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.

Capítulo 7

Experimentación y análisis de resultados

Este capítulo explica detalladamente en qué consistió la experimentación, es decir, cómo se llevó a cabo la identificación de los patrones de consumo de los venezolanos mediante la aplicación de las técnicas seleccionadas para este fin. Primeramente, se explican las corridas que se hicieron para realizar el Análisis de Componentes Principales y la interpretación de los resultados obtenidos a través de la aplicación de dicha técnica. Luego, se habla sobre cómo se condujo la aplicación de las Máquinas de Vectores Soporte como técnica de clasificación no lineal para identificar la propensión al gasto de los venezolanos. Posteriormente, se realiza la interpretación de los resultados obtenidos a través de las MVS y, por último se hace una comparación entre los resultados obtenidos mediante las dos técnicas mencionadas.

7.1 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.

7.1.1 Cálculo y selección de componentes principales

La primera corrida que se realizó fue a través de la matriz de varianza-covarianza para los 22 grupos de gastos descritos en el capítulo del preprocesamiento. En la figura 7.1 se puede observar que, por la regla del codo mencionada en el capítulo 4, gráficamente se pudiesen seleccionar cuatro componentes principales, los cuales juntos explican 63,0843% de la variabilidad de los datos originales, lo cual se consideró un resultado no muy satisfactorio por cuanto se incluían muchos componentes principales. Además, la utilización de la matriz de varianza-covarianza permitió dudar un poco de los resultados pues pudo existir una tendencia a dar mayor importancia a aquellos rubros de gastos que tienen valores de sus varianzas bastante elevados en comparación con los demás. Un ejemplo de esto se observó en los gastos financieros (grupo 18), los cuales tenían su varianza bastante elevada en comparación con la mayoría de los demás rubros de gastos establecidos. Más detalles sobre esta corrida se pueden observar en los anexos.

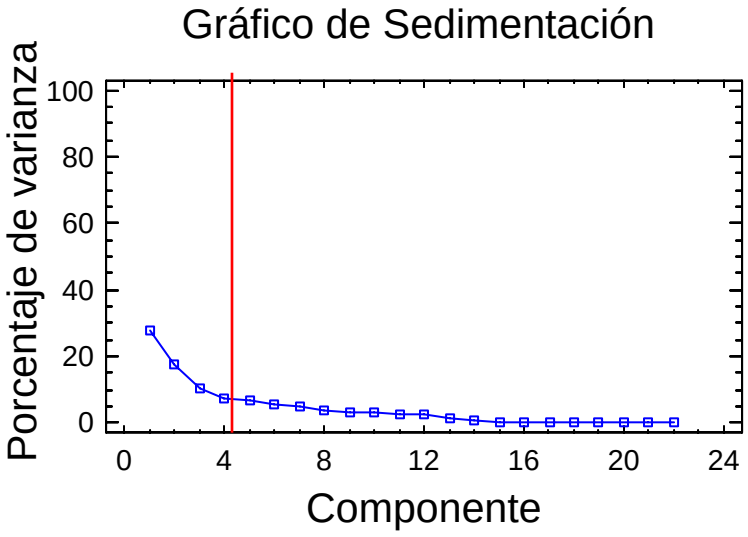


Figura 7.1: Resultados gráficos de la primera corrida del ACP

Debido a que los resultados obtenidos de la primera corrida no se consideraron satisfactorios en cuanto a la dispersión y el número de variables originales, se decidió hacer una reagrupación de los gastos con la idea de tener un menor número de grupos y, de esta manera, los gastos se encontrarían menos dispersos entre tantos grupos. Esta reagrupación se justifica en el hecho de que se observó que existían pocos gastos registrados en algunos de los grupos, lo que llevó a tener 16 nuevos grupos de gastos definitivos, con los cuales se buscaría identificar los patrones de consumo de los venezolanos. La manera como se reagruparon los gastos se explica a continuación:

a) Los primeros ocho grupos permanecieron iguales ya que resultaban bastante representativos y, además, había registros en esos grupos para gran parte de los hogares encuestados. En otras palabras, casi todos los hogares reconocieron haber realizado por lo menos un gasto en cada uno de esos grupos durante los períodos de referencia establecidos. Estos grupos a saber son: 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; y 8) Salud.

b) El nuevo grupo nueve se denomina "Transporte" y resulta de la agrupación de transporte personal y transporte colectivo, los cuales corresponden a los anteriores grupos 9 y 10, respectivamente. El conocer los gastos en los que incurren los venezolanos respecto al transporte en general resulta bastante representativo. Además, se presentaron muchos casos en los que si un hogar incurría en gastos de transporte público no incurría en gastos de transporte personal y viceversa.

c) El naciente grupo diez quedó conformado por los antiguos grupos 11, 12, 13, 14 y 21, los cuales corresponden a telefonía fija, telefonía móvil, Internet, otras comunicaciones y, por último, satélite y cable para TV, respectivamente. A pesar de que inicialmente se quiso observar el comportamiento de los venezolanos respecto a estos

gastos por separado, resultó impertinente continuar estudiándolos de esa manera debido, principalmente, a que se registraron muy pocos desembolsos en cada uno de los grupos. Aproximadamente el 84% de los encuestados señaló no haber incurrido en gastos relacionados con telefonía móvil y casi el 99% no registró gastos en Internet. Casos similares se presentaron con las otras tecnologías de comunicación mencionadas anteriormente. Sin embargo, se unieron todos estos gastos en uno sólo denominado "Comunicaciones" para observar si juntos reflejaban parte de los patrones de consumo que se estaban estudiando.

d) Los anteriores grupos 15,16 y 17 referentes a Educación y cultura, Recreación y esparcimiento, y Artículos, efectos personales y sus servicios, respectivamente, permanecieron de la misma manera pues resultaba interesante estudiarlos por separado. Además, existían suficientes observaciones (más del 50%) de gastos en cada uno de esos grupos. Sin embargo, debido a la nueva enumeración, éstos corresponderían a los nuevos grupos 11, 12 y 13, respectivamente.

e) Los gastos financieros (grupo 18) y los gastos tributarios y legales (grupo 19) pasaron a conformar un único grupo 14 denominado "Gastos financieros, tributarios y legales". Esto debido a que, por separado, ambos tenían muy pocas observaciones y juntos resultaban suficientemente representativos.

f) Los últimos nuevos grupos 15 y 16 denominados "Viajes" y "Otros gastos", respectivamente, son equivalentes a anteriores grupos con los mismos nombres. Aunque un gran número de encuestados no registró desembolsos relacionados con viajes, se decidió no unir este rubro con otro ya que los valores registrados tendían a hacer altos y resultaba interesante destacar cómo era el comportamiento de los venezolanos respecto a gastos en viajes.

Luego de tener la reagrupación de los gastos, se procedió a aplicar el análisis de componentes principales, con *Statgraphics,* a las nuevas variables establecidas. 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. Los resultados detallados de estas corridas se muestran en el anexo II.

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 7.1 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 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 7.2 se puede observar los pesos de los autovalores de todos los componentes y la recta que indica el número de componentes principales a seleccionar. Además, en los anexos se presenta de forma más detallada los resultados obtenidos de esta corrida.

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 7.1: Resultados de la corrida final del ACP

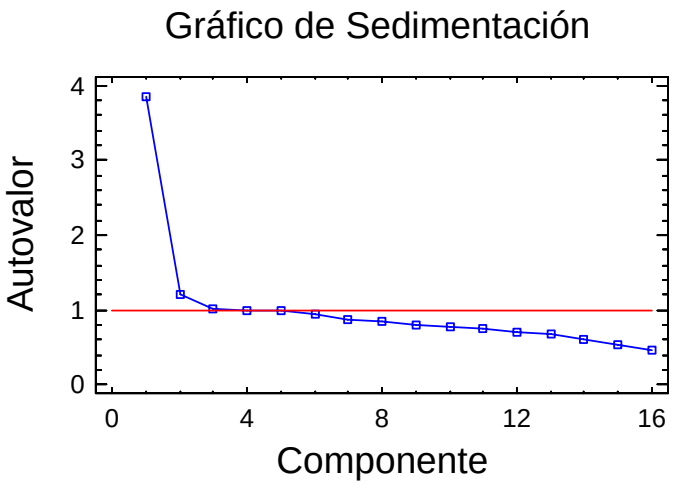


Figura 7.2: Resultados gráficos de la corrida final del ACP

7.1.2 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 7.2.

	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 7.2: Matriz de pesos de los componentes

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.

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. En la figura 7.3 se puede observar el gráfico de pesos para los componentes 1 y 2 de acuerdo a los valores de correlaciones de los grupos de gastos. Encerrados dentro del óvalo azul (línea punteada) se encuentran los gastos que describen el primer eje correspondiente al primer componente principal, mientras que los gastos correspondientes al segundo componente principal, los cuales describen al segundo eje, se encuentran dentro de los óvalos rojos (línea completa).

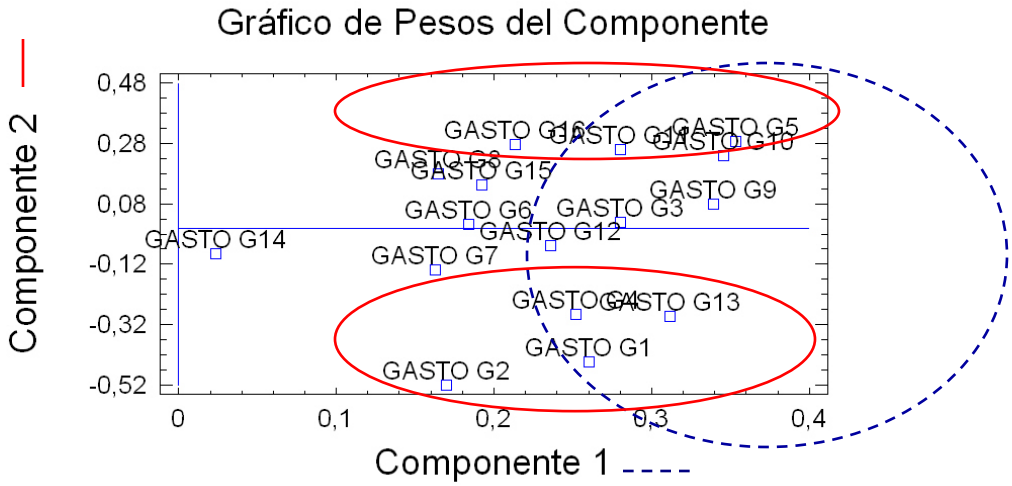


Figura 7.3: Gráfico de pesos de Componente 1 vs. Componente 2

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. En la figura 7.4 se muestra el gráfico de pesos para el tercer y el cuarto componente principal. Dentro de los óvalos azules con línea punteada se encuentran los grupos de gastos que describen el tercer eje, el cual corresponde al tercer componente principal. Asimismo, dentro de los óvalos rojos con línea completa se encuentran los gastos que describen al cuarto eje, definido a partir del cuarto componente principal.

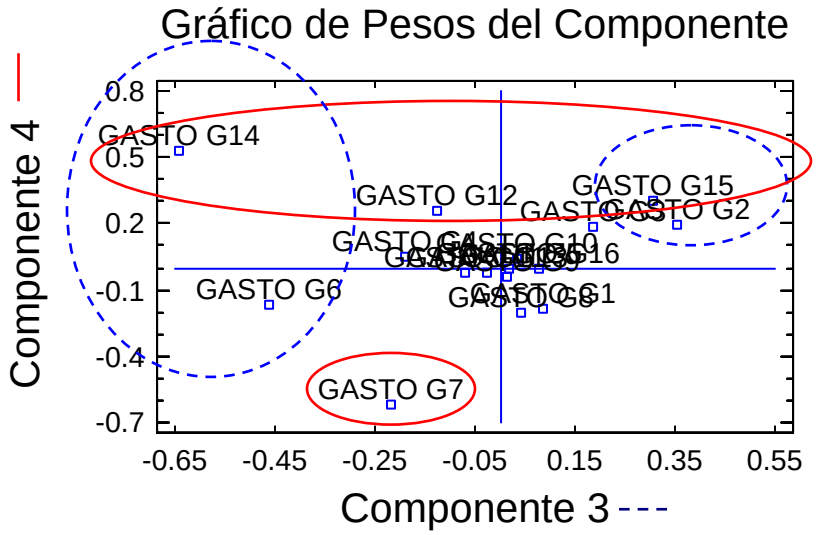


Figura 7.4: Gráfico de pesos de Componente 3 vs. Componente 4

7.1.3 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. Para esto, se recurrió al profesor jubilado Jaime Tinto, quien pertenece al Instituto de Investigaciones Económicas y Sociales de la actual Facultad de Ciencias Económicas y Sociales de la Universidad de Los Andes. El mencionado profesor tiene amplios conocimientos sobre estudios de patrones de consumo de los venezolanos, al igual que ha manejado por muchos años todo lo referente a las encuestas de presupuesto familiar. Su colaboración permitió dar sentido a las nuevas variables obtenidas y nombrarlas adecuadamente, de acuerdo a información referente al ámbito del estudio.

El profesor Tinto determinó necesario eliminar de los componentes las variables que resultaban ambiguas; es decir, sugirió excluir de ciertos componentes aquellas variables que se encontraban dentro de más de uno ellos. 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 7.3, 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.

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.

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 7.3: Interpretación y composición de las variables definidas

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.

7.1.4 Análisis comparativo con resultados de estudios anteriores

Los gastos básicos relacionados con vivienda y sus servicios (alquiler, pago de servicios como agua, luz, gas, etc.) se destacan en el presente estudio como uno de los principales gastos en los que incurren los venezolanos. Incluso este rubro fue el que obtuvo mayor coeficiente de correlación con el primer componente principal. De manera análoga, en el estudio de Garnica (1996), realizado para la ciudad de Mérida a partir de la primera encuestas de presupuesto familiar, se concluyó que el rubro de vivienda y sus servicios fue el aspecto con mayor prioridad en los presupuestos de los merideños para el año 1986. Esta comparación permite inferir que este comportamiento se extiende a todo el territorio nacional y que, a pesar de que el tiempo transcurrido entre una encuesta y otra fue de aproximadamente 12 años, el rubro relacionado con vivienda y sus servicios se mantuvo prevaleciente en el transcurso de esos años.

La investigación de Garnica (1996) sobre ACP en los presupuestos familiares merideños arrojó que los gastos en equipamiento de la vivienda figuraban como una de sus prioridades. Sin embargo, este rubro no se distinguió entre los componentes principales obtenidos en este estudio. Esta diferencia se puede deber a dos situaciones: primera, que el caso que se presenta para el estado Mérida en particular no se extiende para a todo el territorio nacional; segundo, que las prioridades de los venezolanos respecto a este rubro haya cambiado con el paso de los años.

Por otra parte, el ACP aplicado en el estudio de Sosa (2006) permitió conocer la estructura del gasto para el año de 1988, el cual resultó distinto para las ocho regiones estudiadas (Capital, Central, Centro-occidental, Nor-oriental, Los Andes, Guayana, Zulia y Los Llanos). De manera general, la mayoría de las regiones estuvo descrita por las categorías alquiler de vivienda y sus servicios, alimentos y transporte. Opuesto a esto, en el estudio realizado no se pudo estudiar las regiones por separado como inicialmente se pretendió

debido a que las muestras de las diferentes entidades no contaban con un factor de expansión. Sin embargo, en ambos estudios pudo observar que los gastos en vivienda y sus servicios, alimentos y transporte se destacan como gastos básicos a los cuales los venezolanos en general dan prioridad en sus consumos.

7.2 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.

7.2.1 Selección de la 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 aleatorio estratificado. 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 siguiente fórmula:

$$n = \frac{N * Z_{\alpha}^{2} * p * (1-p)}{d^{2} * (N-1) + Z_{\alpha}^{2} * p * (1-p)} \qquad (7.1)$$

En donde N es el tamaño de la población o muestra total, Z es el valor de la variable aleatoria estandarizada para $\alpha/2$ correspondiente a un nivel de confianza o seguridad 1-α, p es una idea del valor aproximado de la proporción poblacional y d es el error máximo de

estimación o precisión deseada para el estudio. Esta fórmula se utilizó debido a que se conocía el tamaño total de la muestra original (población) que era de 8406 observaciones y se deseaba saber cuántas de ellas se debían usar para entrenar las MVS. Un valor de Z de 1,96 para una confiabilidad del 95% resulta, de manera general, bastante buena para estimaciones de tamaños muestrales. Asimismo, el valor de *p* depende de la frecuencia que se espera del factor a estudiar y cuando no se tiene conocimiento sobre esto es común utilizar 0,5 para maximizar el tamaño de la muestra.

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 que se utilizarían se determinaron como se indica a continuación:

$$n_1 = \frac{8406 * (1{,}96)^2 * 0{,}05 * 0{,}95}{((0{,}01)^2 * 8405) + 1{,}96^2 * 0{,}05 * 0{,}95} = \frac{1533{,}89}{1{,}0229} = 1499{,}55 \cong 1500 \qquad (7.2)$$

$$n_2 = \frac{8406 * (1{,}96)^2 * 0{,}05 * 0{,}95}{((0{,}03)^2 * 8405) + 1{,}96^2 * 0{,}05 * 0{,}95} = \frac{1533{,}89}{7{,}75} = 197{,}92 \cong 198 \qquad (7.3)$$

Como se mencionó anteriormente, la selección de las muestras se hizo a través de muestreo aleatorio estratificado. Para esto, fue necesario calcular la proporción de observaciones que cada entidad aportaba a la muestra total y, en función de esas proporciones,

extraer, a través de muestreo aleatorio simple, las cantidades necesarias de cada entidad para conformar la muestra de entrenamiento. Las proporciones de la muestra total y la de los dos tamaños de muestras de entrenamiento se pueden observar en la tabla 7.4.

Entidad	Proporción en la muestra total	Proporción para n=1500	Proporción para n=198
1 – Distrito Federal	9,91%	149	20
2 – Anzoátegui	5,82%	87	12
3 – Apure	0,69%	10	1
4 – Maracay	3,78%	57	7
5 – Barinas	3,53%	53	7
6 – Bolívar	10,79%	162	21
7 – Carabobo	7,79%	117	15
8 – Cojedes	1,21%	18	2
9 – Falcón	1,36%	20	3
10 – Guárico	1,17%	18	2
11 – Lara	5,52%	83	11
12 – Mérida	3,22%	48	6
13 – Miranda	14,10%	212	28
14 – Monagas	1,05%	16	2
15 – Nva. Esparta	1,03%	15	2
16 – Portuguesa	1,56%	23	3
17 – Sucre	1,62%	24	3
18 – Táchira	6,10%	92	12
19 – Trujillo	3,37%	51	7
20 – Yaracuy	0,82%	12	2
21 – Zulia	13,66%	205	27
22 – Amazonas	0,87%	13	2
23 – D. Amacuro	1,03%	15	2

Tabla 7.4: Proporción de cada entidad para muestreo estratificado

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. Más detalles sobre el programa se pueden observar en el anexo III.

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. Sin embargo, esta opción no resulta muy útil cuando se quiere estudiar el desempeño de las máquinas respecto a instancias desconocidas, ya que las mismas ya conocerán todas las instancias con la que se probarán y el error tenderá a cero.

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. Esta opción resulta buena cuando no se tienen estratos que se desean incluir en las muestras, ya que las mismas se extraen de manera aleatoria simple, es decir sin considerar ciertos criterios.

La tercera opción es la utilización de una muestra de prueba distinta a la de entrenamiento que debe suministrarse al programa, la cual es la que se debe usar cuando se desea hacer pruebas respecto a muestras extraídas de manera manual o de alguna manera distinta a las que proporciona el *software.* Esta opción fue la que se utilizó para las pruebas de las muestras extraídas a partir del muestreo aleatorio estratificado que se mencionó previamente.

La cuarta y última opción que brinda el programa mencionado es la conocida como validación cruzada, para la cual se debe especificar el número k de pliegues en los cuales se quiere dividir la muestra total. Entonces, la misma consiste en realizar el entrenamiento con $k-1$ pliegues y la prueba con el pliegue restante. Esto se hace k veces para todas las posibles combinaciones y el error de clasificación se obtendrá mediante el promedio de las k pruebas.

Esta opción ha resultado bastante buena en estudios anteriores y es altamente recomendada por varios autores como Hsu *et al.* (2007). Sin embargo, la composición de los *k* pliegues se hace a través de muestreo aleatorio simple, por lo que se prescinde de la posibilidad de componer las muestras de acuerdo a ciertos criterios como el caso de muestreo aleatorio estratificado.

7.2.2 Transformación de los datos al formato del *software* utilizado

Luego de haber seleccionado el *software* a utilizar y de haber extraído las muestras que se van a utilizar para el entrenamiento, prueba y estudio de generalización de las Máquinas de Vectores Soporte, resulta imprescindible transformar dichas muestras al formato compatible con el *software.* En el caso del estudio que se presenta, el programa seleccionado fue *Weka* en su versión 3.4.12, la cual es una implementación en Java de varios algoritmos de clasificación, regresión, clustering, asociación y visualización. *Weka* tiene grandes ventajas entre las cuales se encuentra que es de fácil implementación debido a su interfaz gráfica y es de libre distribución y difusión. Además, *Weka* "es independiente de la arquitectura, ya que funciona en cualquier plataforma sobre la que se haya una máquina virtual Java disponible" (García, 2007, p. 4).

Los archivos que recibe *Weka* son en formato .arff por sus siglas en inglés *Attribute Relation File Format*, el cual es un archivo texto que describe una lista de instancias que comparten un conjunto de atributos. Estos tipos de archivos fueron desarrollados por el Proyecto de Aprendizaje de Máquinas en el Departamento de Ciencias de la Computación de la Universidad de Waikato para su uso con el *software* de aprendizaje de máquina *Weka*. Dichos archivos se conforman de tres partes, a saber: 1) Cabecera, donde se define el nombre de la relación; 2) Declaraciones de atributos, donde se declaran los atributos que componen el archivo junto a su tipo, el cual puede ser real, entero, fecha o categórico; y 3) Datos, donde se declaran todos los datos de la muestra separando con comas los atributos y con saltos de línea las relaciones o instancias de la muestra.

De esta manera, se procedió a transformar todas las muestras previamente seleccionadas de manera manual a través de muestreo aleatorio estratificado del formato en el que se encontraban (.xls) al formato compatible con *Weka*. Asimismo, se transformó la muestra total de los datos de un formato .xls al formato .arff, con el fin de realizar pruebas posteriores a través de validación cruzada. Luego, se pudo continuar con el siguiente paso necesario para lograr el objetivo buscado, que era poder identificar patrones de consumo de

los venezolanos mediante la técnica de aprendizaje automático denominada Máquinas de Vectores Soporte.

7.2.3 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:

$$k(\bar{x}_i, \bar{x}_j) = \exp(-\gamma \left\| \bar{x}_i - \bar{x}_j \right\|^2) \qquad (7.4)$$

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 7.5.

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.

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 7.5: Configuraciones para las MVS

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.

7.2.4 Selección del tipo de prueba para el clasificador

Como se mencionó anteriormente, los resultados de las diferentes corridas dependen, entre otras cosas, de la manera cómo se pruebe el clasificador. Las opciones para probar el desempeño del clasificador son: 1) Usar el conjunto de entrenamiento; 2) Suministrar un conjunto de datos para la prueba distinto del de entrenamiento; 3) Evaluación con validación cruzada; y 4) Usar un porcentaje de los datos suministrado para el entrenamiento y el resto para la prueba. Para el estudio realizado se utilizaron tanto la segunda como la tercera opción. La segunda se utilizó suministrándole al programa una muestra de datos extraída a través de muestreo aleatorio estratificado y para la prueba, diferentes tamaños de muestras compuestas de instancias diferentes a las utilizadas para el entrenamiento. Esto se debió a que resultaba útil e interesante observar el desempeño de las MVS respecto a la clasificación de instancias desconocidas, con lo que se estudió la habilidad que tienen las mismas para generalizar, lo cual es una ventaja que las caracteriza. El uso de diferentes tamaños de muestras de prueba se debió a que se quiso observar si los mismos influían en los porcentajes de instancias clasificadas correctamente. Por otra parte, se utilizó la tercera opción para medir el desempeño de las distintas configuraciones con esta herramienta que brinda *Weka*, la cual es altamente recomendada por muchos autores, entre los cuales se encuentran Hsu *et al.* Además, la utilización de diferentes modos de medir el clasificador permitió comparar los resultados obtenidos.

7.2.5 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.

7.2.6 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 siguiente tabla.

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

Seguidamente, para la muestra total se realizaron las respectivas pruebas utilizando la validación cruzada con un parámetro de 3; es decir, tres pliegues. Los resultados de esta corrida se muestran en la tabla 7.7. Los porcentajes de instancias clasificadas correctamente mediante este modo de prueba también fueron comparados para las 20 configuraciones

propuestas y, además, la mejor configuración se comparó con la obtenida mediante el primer modo de prueba utilizado.

Utilizando validación cruzada para la muestra total			
Configuración	**C**	γ	**Instancias clasificadas correctamente**
1	1	0,01	97,3114%
2	1	0,1	97,3233%
3	1	1,0	97,6326%
4	1	10,0	98,0966%
5	1	100,0	97,7516%
6	10	0,01	97,3352%
7	10	0,1	97,6802%
8	10	1,0	98,4535%
9	10	10,0	98,7509%
10	10	100,0	98,0371%
11	100	0,01	97,6802%
12	100	0,1	98,5368%
13	100	1,0	99,0602%
14	100	10,0	98,8936%
15	100	100,0	98,1918%
16	1000	0,01	98,5011%
17	**1000**	**0,1**	**99,0959%**
18	1000	1,0	99,2505%
19	1000	10,0	98,9888%
20	1000	100,0	98,2631%

Tabla 7.7: Resultados de todas las configuraciones mediante validación cruzada (n = 8406)

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 7.8.

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

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. Por otro lado, cuando se realiza el entrenamiento de las MVS y las pruebas a través de la opción de validación cruzada que ofrece *Weka,* el mejor resultado se obtiene a través de la configuración 18. Para establecer cuál de ellas se determinaría como mejor configuración, se realizó el entrenamiento y la prueba con cada una utilizando una muestra diferente a las utilizadas en las primeras pruebas. Los resultados obtenidos, suministrando el conjunto de

datos para las pruebas, fueron que con la configuración 17 se logra un 98,8416% de instancias clasificadas correctamente, mientras que con la configuración 18 se consigue un 98,4072%. Por otra parte, utilizando la validación cruzada que ofrece el *software* utilizado, se obtiene con la configuración 17 un 98,5333%; mientras que con la configuración 18 se logra un 98,4667. 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.

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 tabla 7.9 se puede observar los resultados de las pruebas realizadas para un tamaño de muestra de entrenamiento de 1500. Por otro lado, en la figura 7.5 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 7.6 se muestra el comportamiento del modelo respecto al porcentaje de instancias desconocidas clasificadas correctamente.

n = 1500	Tamaño de la muestra de prueba	Instancias clasificadas correctamente	Error absoluto medio	Raíz del error cuadrático medio
Muestra 1	500	98,8%	0,2512	0,3117
Muestra 2	500	99,2%	0,2513	0,3136
Muestra 3	500	98,8%	0,2510	0,3127
Muestra 4	1000	98,7%	0,2512	0,3112
Muestra 5	1000	99,3%	0,2508	0,3123
Muestra 6	1000	99,3%	0,2508	0,3123
Muestra 7	2000	99%	0,2509	0,3112
Muestra 8	2000	99%	0,2512	0,313
Muestra 9	2000	98,9%	0,2512	0,313
Muestra 10	4000	98,95%	0,2509	0,3111
Muestra 11	4000	98,95%	0,2513	0,3132
Muestra 12	4000	99,025%	0,2510	0,3127
Muestra 13	6906	98,885%	0,2510	0,3113
Muestra 14	6906	98,8705%	0,2513	0,3132
Muestra 15	6906	98,9285%	0,2511	0,3128
VARIANZA	-	3,07529E-6	3,17143E-8	7,16952E-7

Tabla 7.9: Resultados de pruebas de la mejor configuración (n = 1500)

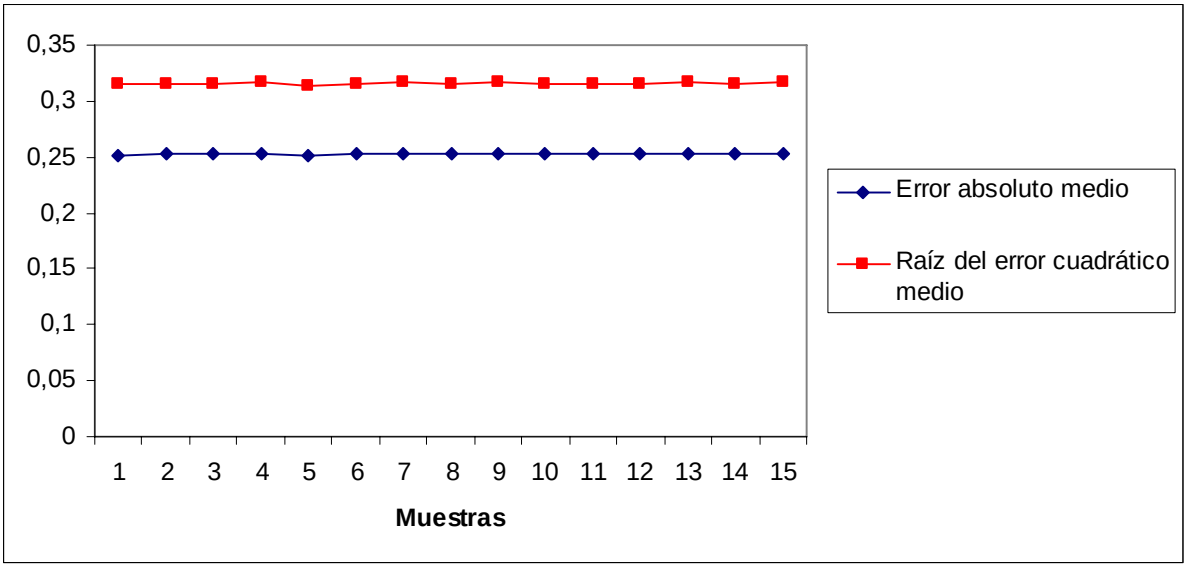


Figura 7.5: Errores para las pruebas de la mejor configuración (n = 1500)

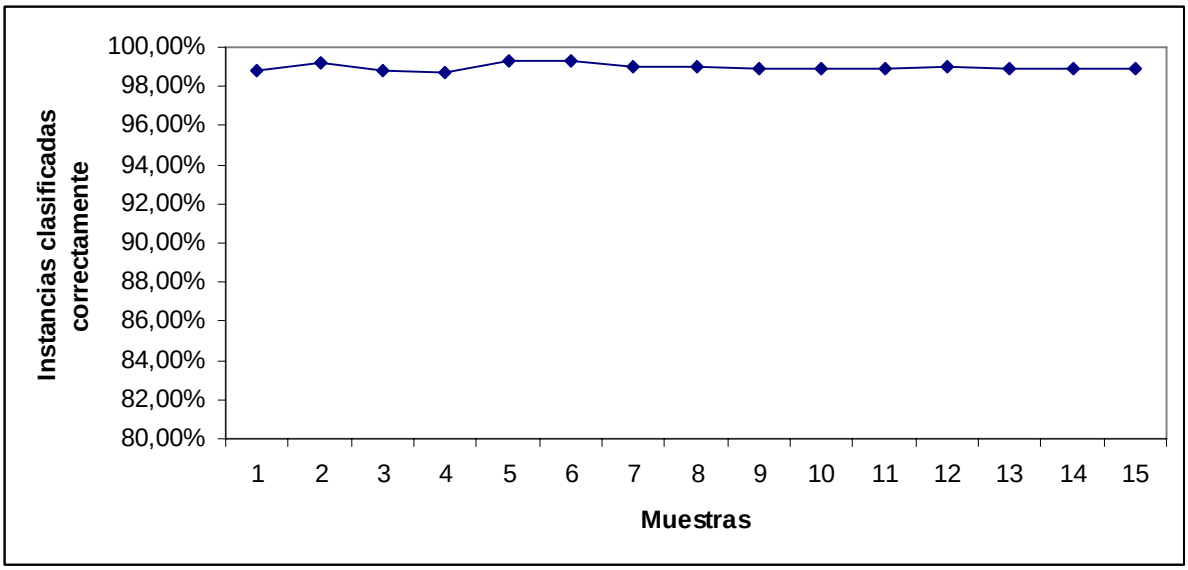


Figura 7.6: Resultados de clasificación para las pruebas de la mejor configuración (n = 1500)

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 resultados obtenidos se muestran en la tabla 7.10.

n = 198	Tamaño de la muestra de prueba	Instancias clasificadas correctamente	Error absoluto medio	Raíz del error cuadrático medio
Muestra 1	500	97,8%	0,2520	0,3148
Muestra 2	500	97,2%	0,2530	0,3162
Muestra 3	500	98,2%	0,2527	0,3161
Muestra 4	1000	97,7%	0,2529	0,3163
Muestra 5	1000	98,9%	0,2513	0,3136
Muestra 6	1000	98,6%	0,2522	0,3153
Muestra 7	2000	97,55%	0,2530	0,3166
Muestra 8	2000	98,2%	0,2524	0,3155
Muestra 9	2000	98%	0,2532	0,3170
Muestra 10	4000	97,725%	0,2528	0,3162
Muestra 11	4000	98,0803%	0,2523	0,3154
Muestra 12	4000	98,2548%	0,2525	0,3158
Muestra 13	8209	97,6976%	0,2529	0,3163
Muestra 14	8209	98,0631%	0,2524	0,3153
Muestra 15	8209	98,0631%	0,253	0,3165
VARIANZA	-	1,76345E-5	2,4638E-7	7,19238E-7

Tabla 7.10: Resultados de pruebas de la mejor configuración (n = 198)

Para una mejor interpretación de los resultados, los valores obtenidos de las distintas pruebas fueron graficados y, posteriormente, analizados. En la figura 7.7 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 7.8 se muestra el comportamiento del modelo seleccionado, para muestras de entrenamiento de 198 ejemplos, respecto al porcentaje de instancias desconocidas clasificadas correctamente.

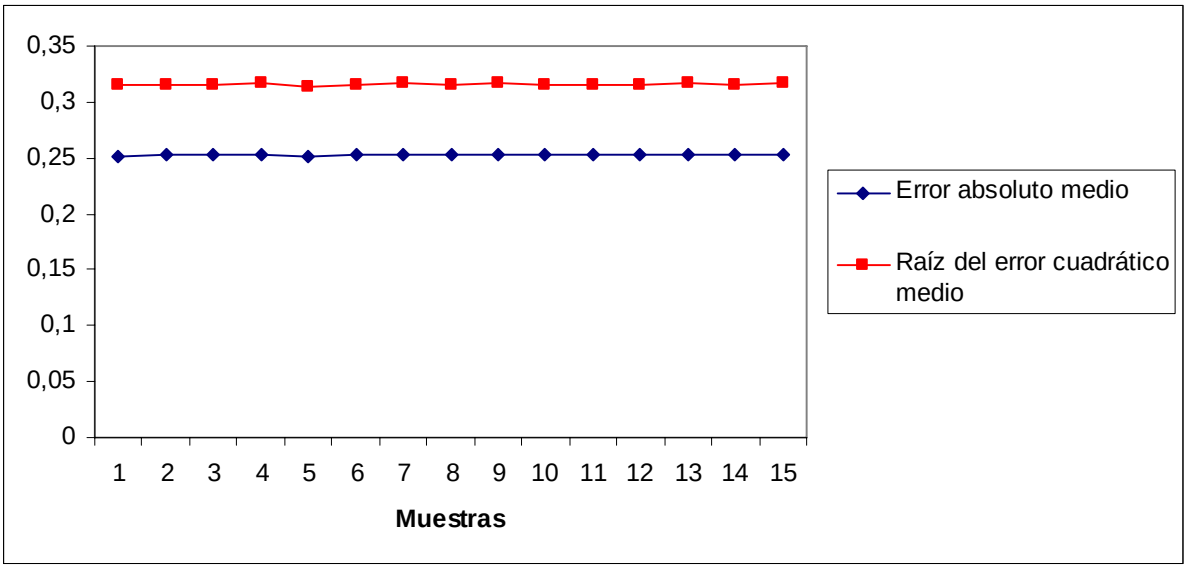


Figura 7.7: Errores para las pruebas de la mejor configuración (n = 198)

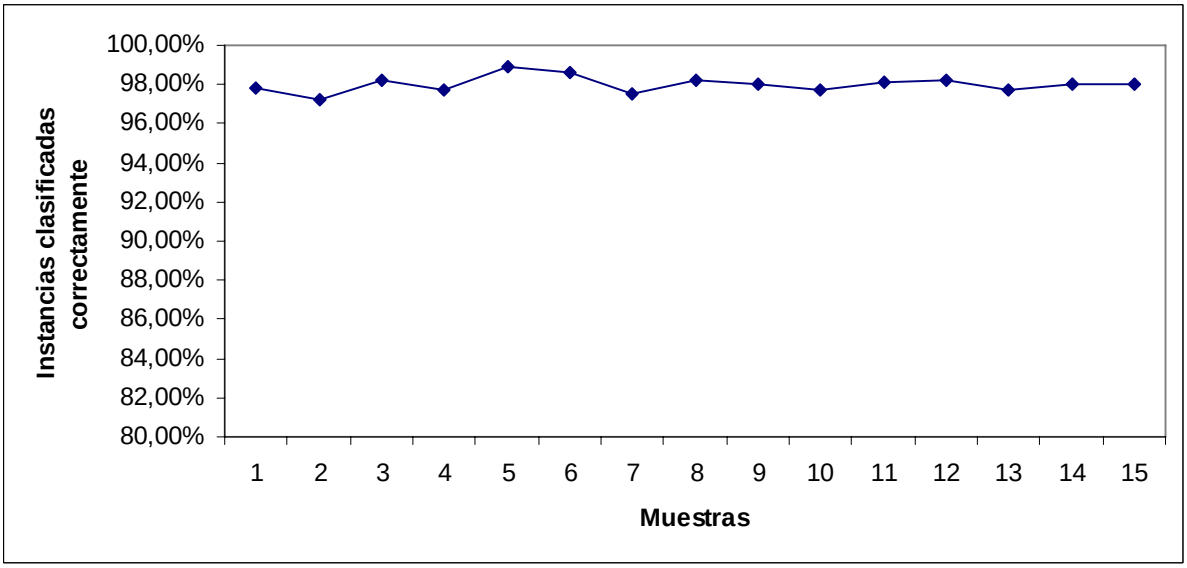


Figura 7.8: Resultados de clasificación para las pruebas de la mejor configuración (n = 198)

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.

7.3 Análisis comparativo de los resultados del ACP y de las MVS

Aunque numéricamente no existen valores que se puedan utilizar para comparar las dos técnicas utilizadas, se puede comparar a ambas de manera analítica respecto a la función que cumplieron cada una de ellas en el estudio y los resultados que se lograron obtener. El Análisis de Componentes Principales, el cual fue utilizado solamente con fines exploratorios, permitió definir las clases que fueron utilizadas para la clasificación con las MVS. Sin embargo, si se quisiera observar el comportamiento de una familia específica con los resultados obtenidos del ACP, resultaría sumamente difícil y hasta imposible en algunos casos, ya que para esto se tendría que recurrir a un gráfico de dispersión en el cual no es fácil detectar la ubicación de dicha familia. Esto se debe a que la mayoría de las familias estudiadas tienden a aglomerarse de acuerdo a su patrón de consumo. Un ejemplo de esto se puede observar en la figura 7.9, la cual muestra el gráfico de dispersión de la última corrida realizada en este estudio.

Por otro lado, si se quisiera identificar el patrón de consumo de una determinada familia utilizando MVS, lo que se tendría que hacer es introducirle esa instancia al modelo seleccionado para poder conocer su modalidad de consumo. Particularmente, el *software* seleccionado permite observar la generalización de todas las instancias de cualquier conjunto de instancias, indicando para cada una la clase a la que pertenece y la clase dentro de la cual se clasificó. Este hecho pone en gran ventaja a las MVS frente a la técnica multivariante ACP. Un ejemplo de esto se puede observar en la figura 7.10 que muestra un ejemplo de los resultados obtenidos a través de *Weka* para la búsqueda de identificación de modalidad de consumo de

cuatro familias seleccionadas de manera aleatoria (diferentes a las usadas para el entrenamiento) que tienen distribuciones de gastos de acuerdo a las cuatro clases definidas.

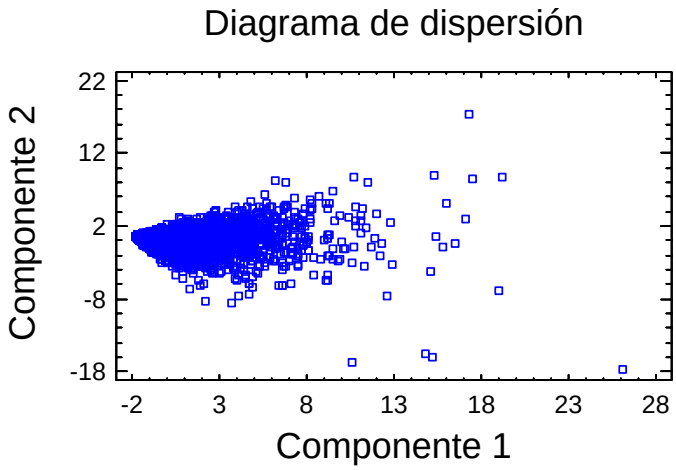


Figura 7.9: Diagrama de dispersión de la última corrida del ACP

```
 === Predictions on test set ===
inst#,    actual, predicted, error, probability distribution
1   4:clase4   4:clase4          0.333  0      0.167 *0.5
2   1:clase1   1:clase1         *0.5    0      0.167  0.333
3   2:clase2   2:clase2          0.333 *0.5    0.167  0
4   3:clase3   1:clase1      +  *0.5    0      0.333  0.167
```

Figura 7.10: Ejemplo de salida de clasificación de MVS con *Weka*

Además, es importante destacar que el *kernel* utilizado para construir el modelo, mapea de manera no lineal los datos a un espacio de características de mayor dimensión para lograr su clasificación. La utilización de las MVS como técnica de clasificación no lineal añade otra ventaja al modelo revelado ya que la mayoría de los problemas de la vida real contienen componentes no lineales envueltos en ellos. Sin embargo, en el caso particular de este estudio, las clases utilizadas para definir la variable de salida de las MVS se extrajeron a partir de una técnica lineal. No obstante, se puede afirmar que las Máquinas de Vectores Soporte se desempeñan excelentemente en el ámbito de clasificación, en general, y en estudios sobre identificación de patrones de consumo, en particular.

Capítulo 8

Conclusiones y recomendaciones

8.1 Conclusiones

- El preprocesamiento de los datos fue una 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 los mismos 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 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 permite 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 herramienta Máquinas de Vectores Soporte arrojó muy buenos resultados en la identificación de patrones de consumo de los venezolanos ya que, usada como método de clasificación no lineal, proporcionó un modelo altamente consistente con una varianza del error absoluto medio mínima (casi nula).

- Las Máquinas de Vectores Soporte mostraron grandes capacidades de generalización ya que con el modelo conseguido, incluso cuando el entrenamiento se hizo a través de menos del 5% de la muestra total de los datos, el error absoluto para las distintas muestras probadas se aproximaba a 0,25 y las instancias desconocidas clasificadas correctamente oscilaban cerca del error 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.

8.2 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.

Bibliografía

Anido, D. (1998), *Sistema Lineal del Gasto: Especificación y Estimación para la Ciudad de Mérida, 1986,* Universidad de Los Andes, Mérida, Venezuela.

Anido D., Orlandoni G. y Quintero, M. (2005), 'Estudio del consumo a partir de las Encuestas Nacionales de Presupuestos Familiares, 1967-2005. El caso de la ciudad de Mérida (Venezuela) ', *AGRAOLIMENTARIA,* Nº 20, pp. 15-41, Universidad de Los Andes, Mérida, Venezuela.

Arellano, R. (2002), *Comportamiento del consumidor. Enfoque América Latina,* McGraw Hill, México.

Carreras, X., Márquez L. y Romero, E. (2004), 'Máquinas de Vectores Soporte', en Hernandez, J., Ramírez, M. y Ferri, C., *Introducción a la Minería de Datos,* pp. 353-382, Editorial Pearson, España.

Chiavenato, I. (2000), *Administración de Recursos Humanos,* 5ta. Edición, McGraw Hill, Santafé de Bogotá, Colombia.

Cristianini, N. y Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods,* Cambridge University Press, United Kingdom.

Díaz, L. (2002), *Estadística multivariada: inferencia y métodos*, Primera edición, Universidad Nacional de Colombia, Bogotá, Colombia.

García, D. (2007), *Manual de Weka,* [Documento WWW], recuperado: http://metaemotion.com/diego.garcia.morate/download/weka.pdf

García, N. (Coordinador) (2006), *Culturas en globalización*, Editorial Nueva Sociedad, Caracas, Venezuela.

García, J., Rodríguez, J. y Vidal J. (2005), *Aprenda Matlab 7.0 como si estuviera en Primero*, Universidad Politécnica de Madrid, España.

Garnica, E. (1996), 'Análisis de componentes principales en los presupuestos familiares', *Revista Economía: Nueva Etapa*, Nº 11, pp. 28-43, Universidad de Los Andes, Mérida, Venezuela.

Giordani, J. (2004), *Presupuesto Familiar*, [Documento WWW], recuperado: http://www.voltairenet.org/article122486.html

Gonzales, R. (2002) 'Reflexiones sobre el consumo: más allá de lo privado y más acá de la condena', *Revista Proposiciones,*Nº 34, pp. 46-71, Santiago de Chile.

Gunn, S. (1998), *Support Vector Machines for classification and regression,* University of Southampton, United Kingdom.

Hsu, Ch., Chang Ch. y Lin Ch. (2007) *A practical guide to Support Vector Classification*, [Documento WWW], recuperado: http://www.csie.ntu.edu.tw/~cjlin

Johnson, D. (2000), *Métodos multivariados aplicados al análisis de datos,* International Thomson Editores, México.

Kotler, P. (1989), *Mercadotecnia,* Editorial 2da. Edición, Prentice Hall, México.

London, D. y Della, A. (1995). *Comportamiento del consumidor. Conceptos y aplicaciones,* 4ta. Edición. McGraw Hill, México.

Lozano, J. (2008), *Introducción al Statgraphics*, [Documento WWW], recuperado: http://www.utadeo.edu.co/comunidades/estudiantes/ciencias_basicas/estadistica/guia_statgraphics.pdf

Marín, J. (2007), *Análisis Multivariante*, [Documento WWW], recuperado: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/

Márquez P., Víctor E. (2002), *Análisis estadístico de los presupuestos familiares en Venezuela*, Instituto de Estadística Aplicada, ULA. Mérida, Venezuela.

Meza, Otilia (2001), 'Notas metodológicas sobre la II Encuesta Nacional de Presupuestos Familiares 1997/1998', *Cuadernos BCV,* Nº 15, pp. 6-52, Venezuela.

Minear, P., Engel, J. y Blackwell, R. (2000), *Consumer Behavior*, 8va. Edición, Hart Court Collage Publishers, Fort Woth.

Moro, Q. y Hurtado, A. (2006), *Introducción al diseño de experimentos para el reconocimiento de patrones, Máquinas de Vectores Soporte*, [Documento WWW], recuperado: http://www.infor.uva.es/~isaac/doctorado/Cap07_SVM.pdf

Pla, L. (1986), *Análisis Multivariado: Método de Componentes Principales,* Secretaría General de la Organización de los Estados Américanos. Washington, D.C.

Platt, J. (2003), *Sequential Minimal Optimization*, [Documento WWW], recuperado: http://research.microsoft.com/~jplatt/smo.html

Sevilla, V. (2006), *El consumo y el ahorro,* [Documento WWW], recuperado: http://www.monografias.com/trabajos14/consumoahorro/consumoahorro.html

Schiffman, L. y Lazar, L. (2001), *Comportamiento del consumidor*, Prentice Hall, México.

Sira-Ramírez, H., Márquez, R., Rivas-Echeverría, F. and Llanes-Santiago, O. (2005), *Control de Sistemas No Lineales*, Pearson Prentice Hall, Madrid, España.

Solomon, M. (1997), *Comportamiento del consumidor*, 3era. Edición, Prentice Hall, México.

Sosa M. (2006), *Estructura del gasto en el presupuesto familiar venezolano Año 1988*, Instituto de Investigaciones Económicas y Sociales, ULA. Mérida, Venezuela

Stanton, W., Etzel, M. y Walter, B. (2004), *Fundamentos del Marketing*, 13ava. Edición, McGraw Hill, México.

Wilkie, W. (1990), *Consumer Behavior*, John Wiley & Sons, 2da. Edición, New York.

Anexo I

Composición de los grupos de gastos definidos

COMPOSICIÓN DE LOS GRUPOS DE LOS 22 GRUPOS DE GASTOS INICIALES

GRUPO 1: ALIMENTOS Y BEBIDAS NO ALCOHÓLICAS

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
10101-10122	Subgrupo Cereales y productos derivados
10201-10233	Subgrupo Carnes de ganado vacuno
10301-10313	Subgrupo Carne de aves
10401-10413	Subgrupo Carne de cerdo
10501-10511	Subgrupo Otras Carnes
10601-10631	Subgrupo Preparados de Carnes
10701-10773	Subgrupo Pescados, mariscos y crustáceos
10801-10835	Subgrupo Leche, queso y huevo
10901-10913	Subgrupo Grasas y aceites Comestibles
11001-11057	Subgrupo Frutas
11101-11137	Subgrupo Hortalizas
11201-11210	Subgrupo Raíces feculantes y derivados
11301-11332	Subgrupo Semillas oleaginosas...
11401-11405	Subgrupo Azúcar
11501-11510	Subgrupo Café, té y cacao
11601-11670	Subgrupo Productos alimenticios varios
11701-11704	Subgrupo Alimentos para niños
11801-11816	Subgrupo Bebidas no alcohólicas
12401-12430	Subgrupo Comidas elaboradas

GRUPO 2: BEBIDAS ALCOHÓLICAS Y TABACO

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
11901-11915	Subgrupo Bebidas Alcohólicas
12201	Subgrupo Bebidas Alcohólicas tomadas fuera del hogar
12301-12306	Subgrupo Tabaco

GRUPO 3: ALIMENTOS Y BEBIDAS TOMADAS FUERA DEL HOGAR

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
12001-12005	Subgrupo Alimentos tomados fuera del hogar
12101	Subgrupo Bebidas No Alcohólicas tomados fuera del hogar

GRUPO 4: VESTIDO Y CALZADO

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
20101-20127	Subgrupo Ropa para Caballeros
20201-20238	Subgrupo Ropa para Damas
20301-20330	Subgrupo Ropa para Niños
20401-20437	Subgrupo Ropa para Niñas
20501-20532	Subgrupo Ropa para Bebé
20601-20653	Subgrupo Accesorios
20701-20729	Subgrupo Otros Artículos Personales
20801-20813	Subgrupo Servicio de confección, reparación y alquiler de prendas de vestir
20901-20911	Subgrupo Calzado para Caballeros
21001-21011	Subgrupo Calzado para Damas
21101-21111	Subgrupo Calzado para Niños
21201-21211	Subgrupo Calzado para Niñas
21301-21304	Subgrupo Calzado para Bebé
21401-21410	Subgrupo Vestimenta, calzado y accesorios de trabajo
21501-21513	Subgrupo Servicio de confección y reparación de calzado
21601-21624	Subgrupo Materiales para la confección o reparación de prendas..

GRUPO 5: VIVIENDA Y SUS SERVICIOS

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
30101-30106	Subgrupo Alquiler de vivienda y servicios
30201-30234	Subgrupo Servicios
30301-30304	Subgrupo Otros combustibles
40801-40108	Parte del subgrupo Servicios Domésticos
90405, 06, 08	Parte del subgrupo Seguros y previsión social

GRUPO 6: MOBILIARIO Y EQUIPOS DEL HOGAR

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
40101-40189	Subgrupo Muebles
40201-40240	Subgrupo Ropa y enseres del hogar
40301-40335	Subgrupo Electrodomésticos
40401-40418	Subgrupo Otros Equipos del Hogar
40501-40550	Subgrupo Utensilios del Hogar

GRUPO 7: MANTENIMIENTO DE LA VIVIENDA

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
30401-30403	Subgrupo Mantenimiento de la vivienda
40601-40662	Subgrupo Gastos diversos del hogar
40701-40723	Subgrupo Servicios para el hogar excepto servicios domésticos
41001-41012	Subgrupo Materiales y repuestos para reparación de mobiliario, enseres y equipos del hogar

GRUPO 8: SALUD

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
50101-50128	Subgrupo Medicinas en General
50201-50239	Subgrupo Artículos Médicos
50301-50329	Subgrupo Salud y otros servicios conexos
50401-50403	Subgrupo Material y repuestos para la reparación de artículos médicos
90409,12,13	Parte del subgrupo Seguros y previsión social

GRUPO 9: TRANSPORTE PERSONAL

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60101-60113	Subgrupo Vehículos de transporte personal
60201-60211	Subgrupo Cauchos, piezas y accesorios para vehículos
60301-60311	Subgrupo Combustible y lubricante para vehículo
60601-60616	Subgrupo Servicios y reparación de vehículos
60701-60705	Subgrupo Otros gastos por vehículos
60901-60909	Subgrupo Materiales y repuestos para la reparación de vehículos
90407,11,14,16,17,21	Parte del subgrupo Seguros y previsión social

GRUPO 10: TRANSPORTE COLECTIVO

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60501,03-07,17,20,22,24	Parte del subgrupo Servicios de Transporte

GRUPO 11: TELEFONÍA FIJA

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60803,06,10-13,22,25,29	Parte del subgrupo Equipos y Servicios de Comunicaciones
70330	Bloqueador de teléfono (aparato)

GRUPO 12: TELEFONÍA MÓVIL

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60814,18,20,23,24,26,27	Parte del subgrupo Equipos y Servicios de Comunicaciones

GRUPO 13: INTERNET

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60815, 60817	Parte del subgrupo Equipos y Servicios de Comunicaciones

GRUPO 14: OTRAS COMUNICACIONES

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60802,04,05,07-09,16	Parte del subgrupo Equipos y Servicios de Comunicaciones

GRUPO 15: EDUCACIÓN Y CULTURA

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
70107,08,15-17,20-22 24-26,29,31,33,34,36 39,45-47,49-52,54,56 57 59-65,67,68,70-72,79,83-85,87	Parte del subgrupo Artículos, equipos y accesorios relacionados con la educación y esparcimiento
70301,08-14,17,23,34	Parte del subgrupo Otros Equipos para educación y esparcimiento
70421-34,37,47,49,56 57,63,65,66 68,69,73-75	Parte del subgrupo Servicios relacionados con educación y esparcimiento
70501,04,07	Parte del subgrupo Material para elaboración y reparación de equipos de educación, esparcimiento y deporte
90249	Parte del subgrupo Servicios y otros pagos n.e.o.p.

GRUPO 16: RECREACIÓN Y ESPARCIMIENTO

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
70101-06,09-14,18,19,23,27 28,30 32,35,37,38 40-44,48,53 55,66,69,73-78,80-82 86	Parte del subgrupo Artículos, equipos y accesorios relacionados con la educación y esparcimiento
70201-70238	Subgrupo Equipos y Artículos Deportivos
70302-07,15,18-22 24-29,32,33	Parte del subgrupo Otros Equipos para educación y esparcimiento
70401-17,19,20,35,36,38,40- 46,48,50-55,58-60,64,67,70-71	Parte del subgrupo Servicios relacionados con educación y esparcimiento
70502,03,05 06,08,09	Parte del subgrupo Material para elaboración y reparación de equipos de educación, esparcimiento y deporte
90252,90256	Parte del subgrupo Servicios y otros pagos n.e.o.p.
90316	Parte del subgrupo Impuestos, gastos legales, tasas y otros

GRUPO 17: ARTÍCULOS, EFECTOS PERSONALES Y SERVICIOS DIVERSOS

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
80101-80168	Subgrupo Artículos y servicios de cuidado personal
80201-80220	Subgrupo Efectos Personales
80301-80302	Subgrupo Materiales para fabric. y repara. de articulos personales

GRUPO 18: GASTOS FINANCIEROS

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
90102-90113	Subgrupo Servicios Financieros

GRUPO 19: GASTOS TRIBUTARIOS Y LEGALES

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
90225,27,28	Parte del subgrupo Servicios y otros pagos
90301-11,17,18,22,23,26,28-36	Parte del subgrupo Impuestos, gastos legales, tasas y otros
90501	Subgrupo Intereses sobre préstamos

GRUPO 20: OTROS GASTOS

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60518,19	Parte del subgrupo Servicios de Transporte
70316,70331	Parte del subgrupo Otros Equipos para educación y esparcimiento
70472	Parte del subgrupo Servicios relacionados con educación, esparc. y deporte
90201-24,26 29-48,50,51 53-55,57	Parte del subgrupo Servicios y otros gastos n.e.o.p.
90320,26	Parte del subgrupo Impuesto, gastos legales, tasas...
90401-04,15,18-20,22	Parte del subgrupo Seguro social y prevision
90601-90602	Subgrupo Otros gastos n.e.o.p.

GRUPO 21: SATÉLITE Y CABLE PARA TV

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60801,19,21,28	Parte del subgrupo Equipos y servicios de comunicaciones

GRUPO 22: VIAJES

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60401-60413	Subgrupo Servicios de alquiler de vehículos de transporte
60502, 08-16,21,23	Parte del subgrupo Servicios de Transporte
70418,61,62	Parte del subgrupo Servicios relacionado con educ., esparcimiento y deporte
90312-15,19,24,25	Parte del subgrupo Impuestos, gastos legales, tasas y otros
90410	Parte del subgrupo Seguros y previsión social

COMPOSICIÓN DE LOS GRUPOS DE LOS 16 GRUPOS DE GASTOS FINALES

GRUPO 1: ALIMENTOS Y BEBIDAS NO ALCOHÓLICAS

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
10101-10122	Subgrupo Cereales y productos derivados
10201-10233	Subgrupo Carnes de ganado vacuno
10301-10313	Subgrupo Carne de aves
10401-10413	Subgrupo Carne de cerdo
10501-10511	Subgrupo Otras Carnes
10601-10631	Subgrupo Preparados de Carnes
10701-10773	Subgrupo Pescados, mariscos y crustáceos
10801-10835	Subgrupo Leche, queso y huevo
10901-10913	Subgrupo Grasas y aceites Comestibles
11001-11057	Subgrupo Frutas
11101-11137	Subgrupo Hortalizas
11201-11210	Subgrupo Raíces feculantes y derivados
11301-11332	Subgrupo Semillas oleaginosas...
11401-11405	Subgrupo Azúcar
11501-11510	Subgrupo Café, té y cacao
11601-11670	Subgrupo Productos alimenticios varios
11701-11704	Subgrupo Alimentos para niños
11801-11816	Subgrupo Bebidas no alcohólicas
12401-12430	Subgrupo Comidas elaboradas

GRUPO 2: BEBIDAS ALCOHÓLICAS Y TABACO

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
11901-11915	Subgrupo Bebidas Alcohólicas
12201	Subgrupo Bebidas Alcohólicas tomadas fuera del hogar
12301-12306	Subgrupo Tabaco

GRUPO 3: ALIMENTOS Y BEBIDAS TOMADAS FUERA DEL HOGAR

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
12001-12005	Subgrupo Alimentos tomados fuera del hogar
12101	Subgrupo Bebidas No Alcohólicas tomados fuera del hogar

GRUPO 4: VESTIDO Y CALZADO

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
20101-20127	Subgrupo Ropa para Caballeros
20201-20238	Subgrupo Ropa para Damas
20301-20330	Subgrupo Ropa para Niños
20401-20437	Subgrupo Ropa para Niñas
20501-20532	Subgrupo Ropa para Bebé
20601-20653	Subgrupo Accesorios
20701-20729	Subgrupo Otros Artículos Personales
20801-20813	Subgrupo Servicio de confección, reparación y alquiler de prendas de vestir
20901-20911	Subgrupo Calzado para Caballeros
21001-21011	Subgrupo Calzado para Damas
21101-21111	Subgrupo Calzado para Niños
21201-21211	Subgrupo Calzado para Niñas
21301-21304	Subgrupo Calzado para Bebé
21401-21410	Subgrupo Vestimenta, calzado y accesorios de trabajo
21501-21513	Subgrupo Servicio de confección y reparación de calzado
21601-21624	Subgrupo Materiales para la confección o reparación de prendas..

GRUPO *5*: VIVIENDA Y SUS SERVICIOS

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
30101-30106	Subgrupo Alquiler de vivienda y servicios
30201-30234	Subgrupo Servicios
30301-30304	Subgrupo Otros combustibles
40801-40108	Parte del subgrupo Servicios Domésticos
90405, 06, 08	Parte del subgrupo Seguros y previsión social

GRUPO 6: MOBILIARIO Y EQUIPOS DEL HOGAR

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
40101-40189	Subgrupo Muebles
40201-40240	Subgrupo Ropa y enseres del hogar
40301-40335	Subgrupo Electrodomésticos
40401-40418	Subgrupo Otros Equipos del Hogar
40501-40550	Subgrupo Utensilios del Hogar

GRUPO 7: MANTENIMIENTO DE LA VIVIENDA

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
30401-30403	Subgrupo Mantenimiento de la vivienda
40601-40662	Subgrupo Gastos diversos del hogar
40701-40723	Subgrupo Servicios para el hogar excepto servicios domésticos
41001-41012	Subgrupo Materiales y repuestos para reparación de mobiliario, enseres y equipos del hogar

GRUPO 8: SALUD

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
50101-50128	Subgrupo Medicinas en General
50201-50239	Subgrupo Artículos Médicos
50301-50329	Subgrupo Salud y otros servicios conexos
50401-50403	Subgrupo Material y repuestos para la reparación de artículos médicos
90409,12,13	Parte del subgrupo Seguros y previsión social

GRUPO 9: TRANSPORTE

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60101-60113	Subgrupo Vehículos de transporte personal
60201-60211	Subgrupo Cauchos, piezas y accesorios para vehículos
60301-60311	Subgrupo Combustible y lubricante para vehículo
60501,03-07,17,20,22,24	Parte del subgrupo Servicios de Transporte
60601-60616	Subgrupo Servicios y reparación de vehículos
60701-60705	Subgrupo Otros gastos por vehículos
60901-60909	Subgrupo Materiales y repuestos para la reparación de vehículos
90407,11,14,16,17,21	Parte del subgrupo Seguros y previsión social

GRUPO 10: COMNUNICACIÓN

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60801-60829	Subgrupo Equipos y Servicios de Comunicaciones
70330	Bloqueador de teléfono (aparato)

GRUPO 11: EDUCACIÓN Y CULTURA

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
70107,08,15-17,20-22 24-26,29,31,33,34,36 39,45-47,49-52,54,56 57 59-65,67,68,70-72,79,83-85,87	Parte del subgrupo Artículos, equipos y accesorios relacionados con la educación y esparcimiento
70301,08-14,17,23,34	Parte del subgrupo Otros Equipos para educación y esparcimiento
70421-34,37,47,49,56 57,63,65,66 68,69,73-75	Parte del subgrupo Servicios relacionados con educación y esparcimiento
70501,04,07	Parte del subgrupo Material para elaboración y reparación de equipos de educ., esparcimiento y deporte
90249	Parte del subgrupo Servicios y otros pagos n.e.o.p.

GRUPO 12: RECREACIÓN Y ESPARCIMIENTO

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
70101-06,09-14,18,19,23,27 28,30 32,35,37,38 40-44,48,53 55,66,69,73-78,80-82 86	Parte del subgrupo Artículos, equipos y accesorios relacionados con la educación y esparcimiento
70201-70238	Subgrupo Equipos y Artículos Deportivos
70302-07,15,18-22 24-29,32,33	Parte del subgrupo Otros Equipos para educación y esparcimiento
70401-17,19,20,35,36,38,40- 46,48,50-55,58-60,64,67,70-71	Parte del subgrupo Servicios relacionados con educación y esparcimiento
70502,03,05 06,08,09	Parte del subgrupo Material para elaboración y reparación de equipos de educación, esparcimiento y deporte
90252,90256	Parte del subgrupo Servicios y otros pagos n.e.o.p.
90316	Parte del subgrupo Impuestos, gastos legales, tasas y otros

GRUPO 13: ARTÍCULOS, EFECTOS PERSONALES Y SERVICIOS DIVERSOS

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
80101-80168	Subgrupo Artículos y servicios de cuidado personal
80201-80220	Subgrupo Efectos Personales
80301-80302	Subgrupo Materiales para fabric. y repara. de art. personales

GRUPO 14: GASTOS FINANCIEROS, TRIBUTARIOS Y LEGALES

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
90102-90113	Subgrupo Servicios Financieros
90225,27,28	Parte del subgrupo Servicios y otros pagos
90301-11,17,18,22,23,26,28-36	Parte del subgrupo Impuestos, gastos legales, tasas y otros
90501	Subgrupo Intereses sobre préstamos

GRUPO 15: VIAJES

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60401-60413	Subgrupo Servicios de alquiler de vehículos de transporte
60502, 08-16,21,23	Parte del subgrupo Servicios de Transporte
70418,61,62	Parte del subgrupo Servicios relacionado con educación, esparcimiento y deporte
90312-15,19,24,25	Parte del subgrupo Impuestos, gastos legales, tasas y otros
90410	Parte del subgrupo Seguros y previsión social

GRUPO 16: OTROS GASTOS

CÓDIGO DEL PRODUCTO	DESCRIPCIÓN DEL GRUPO ORIGINAL
60518,19	Parte del subgrupo Servicios de Transporte
70316,70331	Parte del subgrupo Otros Equipos para educación y esparcimiento
70472	Parte del subgrupo Servicios relacionados con educación, esparcimiento y deporte
90201-24,26 29-48,50,51 53-55,57	Parte del subgrupo Servicios y otros gastos n.e.o.p.
90320,26	Parte del subgrupo Impuesto, gastos legales, tasas...
90401-04,15,18-20,22	Parte del subgrupo Seguro social y prevision
90601-90602	Subgrupo Otros gastos n.e.o.p.

Anexo II

Resultados de corridas del Análisis de Componentes Principales

PRIMERA CORRIDA DEL ACP

MATRIZ DE VARIANZA – COVARIANZA PARA 22 GRUPOS DE GASTOS

DATOS ORIGINALES

Datos/Variables:
GASTO G1, GASTO G2, GASTO G3, GASTO G4, GASTO G5,
GASTO G6, GASTO G7, GASTO G8, GASTO G9, GASTO G10,
GASTO G11, GASTO G12, GASTO G13, GASTO G14, GASTO G15
GASTO G16, GASTO G17, GASTO G18, GASTO G19, GASTO G20
GASTO G21, GASTO G22

Entrada de datos: observaciones
Número de casos completos: 8406
Tratamiento de valor perdido: lista considerada
Estandarizado: no

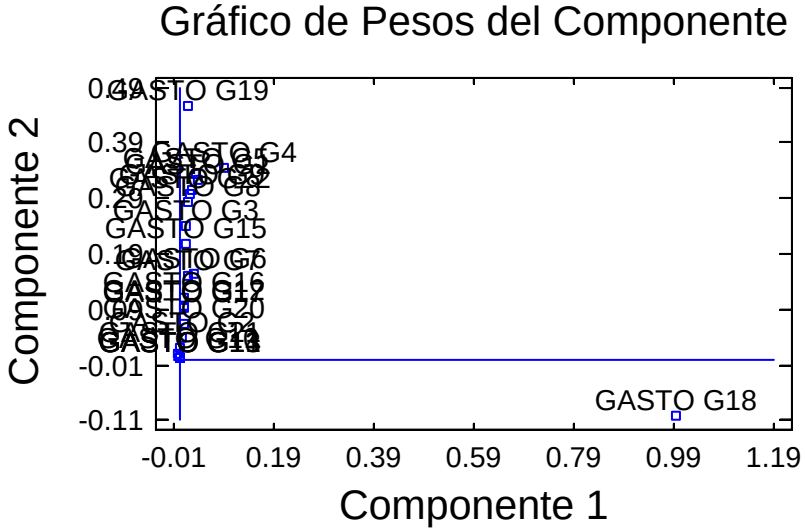
Número de componentes extraídos: 4

Análisis de Componentes Principales

Componente Número	Autovalor	Porcentaje de Varianza	Acumulado Porcentaje
1	1.74576E10	27.942	27.942
2	1.09077E10	17.459	45.401
3	6.56183E9	10.503	55.904
4	4.48636E9	7.181	63.084
5	4.29345E9	6.872	69.956
6	3.60316E9	5.767	75.723
7	3.08214E9	4.933	80.657
8	2.38216E9	3.813	84.469
9	2.15883E9	3.455	87.925
10	2.01623E9	3.227	91.152
11	1.75924E9	2.816	93.968
12	1.45675E9	2.332	96.299
13	8.8287E8	1.413	97.712
14	4.36641E8	0.699	98.411
15	3.05263E8	0.489	98.900
16	2.5928E8	0.415	99.315
17	2.38845E8	0.382	99.697
18	1.40504E8	0.225	99.922
19	3.45453E7	0.055	99.977
20	6.331E6	0.010	99.987
21	5.3695E6	0.009	99.996
22	2.49511E6	0.004	100.000

Tabla de Pesos de los Componentes

	Componente 1	Componente 2	Componente 3	Componente 4
GASTO G1	0.038801	0.324702	-0.243505	-0.27896
GASTO G2	0.00424308	0.0382895	-0.0260678	-0.019565
GASTO G3	0.0131526	0.241765	-0.106368	0.0241773
GASTO G4	0.0910425	0.344593	-0.305495	-0.502485
GASTO G5	0.0351527	0.333512	-0.111939	0.0871075
GASTO G6	0.0298312	0.155431	-0.00824278	-0.0843007
GASTO G7	0.0179779	0.149218	-0.0270423	-0.0927742
GASTO G8	0.0168423	0.282685	-0.181462	0.358561
GASTO G9	0.026712	0.305674	-0.100819	-0.0201434
GASTO G10	-0.0001725	0.00873037	-0.0103071	-0.0174831
GASTO G11	0.0000891982	0.00180242	-0.00154266	-0.0000563572
GASTO G12	0.0100256	0.09521	-0.0290132	0.00802744
GASTO G13	0.000200261	0.00357849	-0.000169918	0.00248104
GASTO G14	0.0000718609	0.00111266	0.000471751	-0.000475333
GASTO G15	0.0139974	0.206896	-0.0753555	0.0104685
GASTO G16	0.0116257	0.109888	-0.0511986	-0.0115916
GASTO G17	0.00801313	0.0917024	-0.0487426	-0.0390186
GASTO G18	0.992597	-0.100681	0.0389295	0.0390601
GASTO G19	0.0172401	0.456386	0.87004	-0.100837
GASTO G20	0.00855816	0.0640187	-0.0221919	0.020475
GASTO G21	0.00289016	0.0223029	-0.00191664	0.00502613
GASTO G22	0.021098	0.299059	-0.0914	0.708639



SEGUNDA CORRIDA

MATRIZ DE VARIANZA – COVARIANZA PARA 16 GRUPOS DE GASTOS

DATOS ORIGINALES

Datos/Variables:
GASTO G1, GASTO G2, GASTO G3, GASTO G4, GASTO G5,
GASTO G6, GASTO G7, GASTO G8, GASTO G9, GASTO G10,
GASTO G11, GASTO G12, GASTO G13, GASTO G14, GASTO G15
GASTO G16

Entrada de datos: observaciones

Número de casos completos: 8406

Tratamiento de valor perdido: lista considerada

Estandarizado: no

Número de componentes extraídos: 4

Análisis de Componentes Principales

Componente Número	Autovalor	Porcentaje de Varianza	Acumulado Porcentaje
1	1.05689E10	21.840	21.840
2	8.52683E9	17.621	39.461
3	5.94443E9	12.284	51.745
4	4.46956E9	9.236	60.981
5	3.73844E9	7.725	68.706
6	3.21197E9	6.637	75.344
7	2.40358E9	4.967	80.311
8	2.18041E9	4.506	84.817
9	1.98448E9	4.101	88.918
10	1.74163E9	3.599	92.517
11	1.4486E9	2.994	95.510
12	8.83052E8	1.825	97.335
13	4.3732E8	0.904	98.239
14	3.59503E8	0.743	98.981
15	2.55064E8	0.527	99.509
16	2.37814E8	0.491	100.000

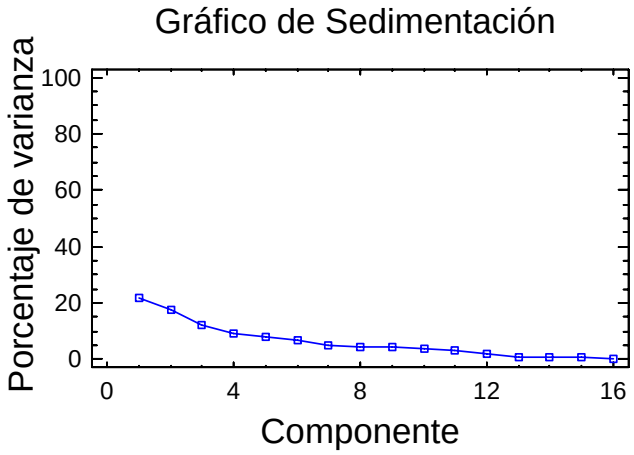
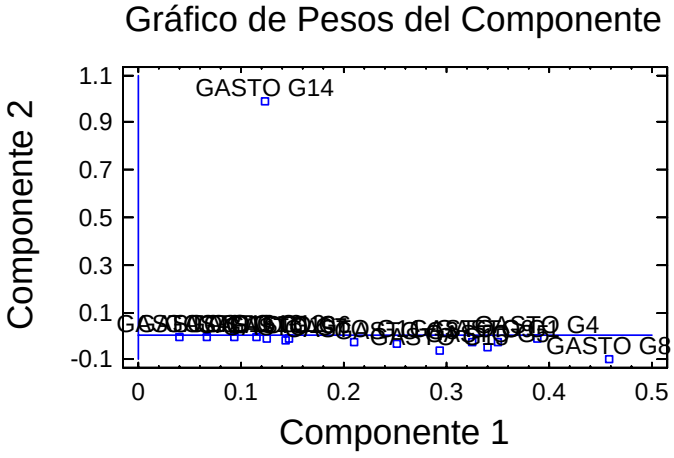


Tabla de Pesos de los Componentes

	Componente 1	Componente 2	Componente 3	Componente 4
GASTO G1	0.34946	-0.0273154	0.173722	0.313056
GASTO G2	0.0405439	-0.00493257	0.0288347	0.0135173
GASTO G3	0.251841	-0.0296303	0.119133	-0.0495344
GASTO G4	0.389043	-0.0137668	0.361161	0.3531
GASTO G5	0.340713	-0.0441107	0.0910367	-0.0894348
GASTO G6	0.147193	-0.0131072	0.0724292	0.0545783
GASTO G7	0.143276	-0.0172869	0.0769789	0.0901635
GASTO G8	0.458448	-0.0971567	-0.865792	0.0483365
GASTO G9	0.325013	-0.0285743	0.115832	0.0515689
GASTO G10	0.12495	-0.0121974	0.0389587	-0.0191076
GASTO G11	0.210001	-0.022664	0.0681899	-0.0135609
GASTO G12	0.114252	-0.00418827	0.0562107	-0.0065852
GASTO G13	0.0926647	-0.00290696	0.045765	0.0251884
GASTO G14	0.123624	0.990219	-0.0482478	-0.0406813
GASTO G15	0.29393	-0.0649449	0.166454	-0.864398
GASTO G16	0.0658539	-0.00708964	0.0207438	-0.022628



TERCERA CORRIDA

MATRIZ DE VARIANZA – COVARIANZA PARA 16 GRUPOS DE GASTOS

DATOS ESCALADOS

Datos/Variables:
GASTO G1, GASTO G2, GASTO G3, GASTO G4, GASTO G5,
GASTO G6, GASTO G7, GASTO G8, GASTO G9, GASTO G10,
GASTO G11, GASTO G12, GASTO G13, GASTO G14, GASTO G15
GASTO G16

Entrada de datos: observaciones

Número de casos completos: 8406

Tratamiento de valor perdido: lista considerada

Estandarizado: no

Número de componentes extraídos: 4

Análisis de Componentes Principales

Componente Número	Autovalor	Porcentaje de Varianza	Acumulado Porcentaje
1	0.0153312	33.690	33.690
2	0.00479769	10.543	44.233
3	0.00387538	8.516	52.750
4	0.00380392	8.359	61.109
5	0.00303047	6.659	67.768
6	0.00261049	5.737	73.505
7	0.00253128	5.563	79.067
8	0.00186276	4.093	83.161
9	0.0013998	3.076	86.237
10	0.00131062	2.880	89.117
11	0.00113562	2.496	91.612
12	0.000991782	2.179	93.792
13	0.000882289	1.939	95.731
14	0.000843843	1.854	97.585
15	0.0006672	1.466	99.051
16	0.000431772	0.949	100.000

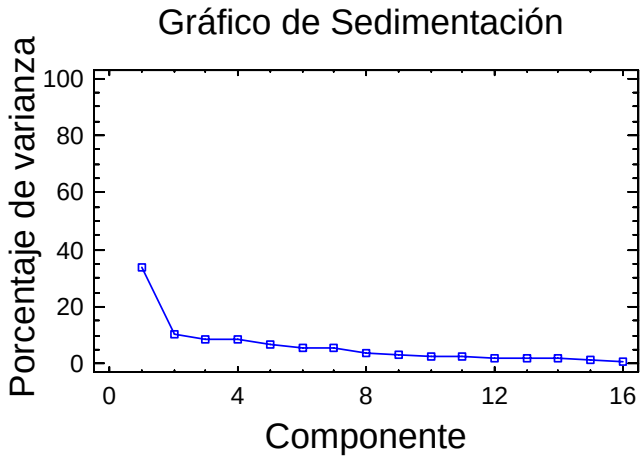
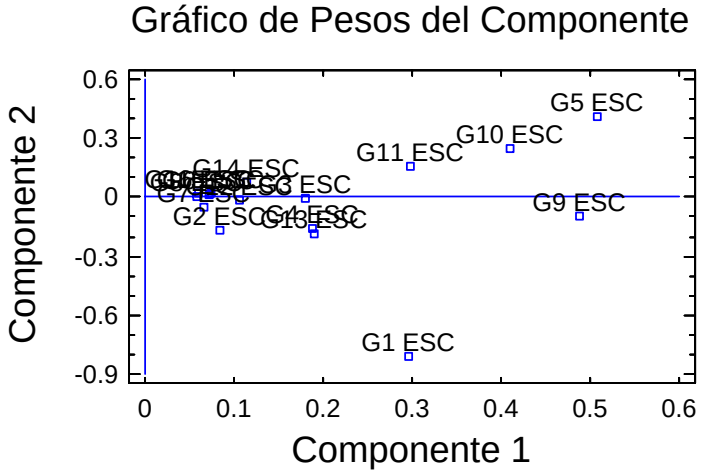


Tabla de Pesos de los Componentes

	Componentes 1	Componentes 2	Componentes 3	Componentes 4
G1 ESC	0.296078	-0.803109	-0.312594	-0.0449
G2 ESC	0.0838473	-0.165495	-0.0203037	-0.00366362
G3 ESC	0.180931	-0.00670407	0.0686737	-0.00649051
G4 ESC	0.187657	-0.155682	-0.041708	0.142257
G5 ESC	0.50896	0.407688	-0.373032	-0.163271
G6 ESC	0.0726148	0.0107114	-0.0198717	0.0565376
G7 ESC	0.0665746	-0.0525885	-0.00578332	0.0265589
G8 ESC	0.0575147	-0.0000773361	-0.00829092	-0.0122229
G9 ESC	0.4885	-0.0923138	0.81023	-0.00130574
G10 ESC	0.410097	0.246932	0.0061029	-0.0224921
G11 ESC	0.298235	0.153321	-0.281062	-0.119622
G12 ESC	0.106792	-0.0145462	-0.0100103	0.0141301
G13 ESC	0.190097	-0.183881	-0.0912873	-0.000402644
G14 ESC	0.113492	0.0792615	-0.103069	0.965348
G15 ESC	0.0743785	0.0243665	-0.0347031	0.00818959
G16 ESC	0.059943	0.0187183	0.000713023	-0.000932442



CUARTA CORRIDA

MATRIZ DE CORRELACIÓN PARA LOS 16 GRUPOS DE GASTOS

DATOS ORIGINALES

Datos/Variables:
GASTO G1, GASTO G2, GASTO G3, GASTO G4, GASTO G5,
GASTO G6, GASTO G7, GASTO G8, GASTO G9, GASTO G10,
GASTO G11, GASTO G12, GASTO G13, GASTO G14, GASTO G15
GASTO 16

Entrada de datos: observaciones

Número de casos completos: 8405

Tratamiento de valor perdido: lista considerada

Estandarizado: si

Número de componentes extraídos: 4

Análisis de Componentes Principales

Componente Número	Autovalor	Porcentaje de Varianza	Acumulado Porcentaje
1	3.85642	24.103	24.103
2	1.1995	7.497	31.600
3	1.02202	6.388	37.987
4	0.997374	6.234	44.221
5	0.981782	6.136	50.357
6	0.932366	5.827	56.184
7	0.876928	5.481	61.665
8	0.843248	5.270	66.935
9	0.796916	4.981	71.916
10	0.783433	4.896	76.812
11	0.751634	4.698	81.510
12	0.702376	4.390	85.900
13	0.668401	4.178	90.078
14	0.603883	3.774	93.852
15	0.528108	3.301	97.152
16	0.455602	2.848	100.000

Anexo III

Programa de Matlab para muestreo aleatorio estratificado

PROGRAMA EN MATLAB PARA EXTRACCIÓN DE MUESTRAS ALEATORIAS SIN REPOSICIÓN

```
function MUESTRAS()
% Función diseñada para tomar dos limites (superior e
% inferior) y luego dado un valor comprendido entre:
% límite _ superior - límite _ inferior, calcular
% aleatoriamente esa cantidad de números y luego
% imprimirlos en pantalla.
% Diseñado por Francisco A. Justo T., Estudiante de Ing. De
% Sistemas

clc
clear all

l_sup = input ('Ingrese el Límite Superior: ')
l_inf = input ('Ingrese el Límite Inferior: ')
num = input ('Ingrese el tamaño de la muestras: ')

cant_muestras = l_sup - l_inf +1;
n = 1;
buscar_vect = 0;
vect = 0;

while n <= num

    otra = floor(rand()*(l_sup - l_inf)+l_inf);
    for j=1:length(vect)
        if otra == vect(j)
            buscar_vect = 1;
        end
    end
    if buscar_vect == 0
        vect(n) = otra;
        n = n+1;
    end
    if buscar_vect == 1
        %i = i-1;
        buscar_vect = 0;
    end

end
disp('Valores XD ')
vect = sort(vect')

[numeric,tex,row]= xlsread('MATRIZ.xls');
n=1;
```

```
for I=1:length(vect)
    for J=1:length(numeric)
       if vect(I)==numeric(J)
           entre(n,:)=numeric(J,:);
           n=n+1;
       end
    end
end

vect2 = l_inf:l_sup;
vect = vect';
Bandera = 0;
for J=1:length(vect)
    cont = length(vect2);
    for I=1:cont
        if vect(J) == vect2(I)
            Bandera = 1;
        end
    end
    if Bandera == 1
        Badenra = 0;
        for K=1:cont
            if K <= cont
             if vect(J) == vect2(K)
                vect2(K)= [];
                cont = cont -1;
             end
            end
        end
    end
end

n=1;
for I=1:length(vect2)
    for J=1:length(numeric)
       if vect2(I)==numeric(J)
           comp(n,:)=numeric(J,:);
           n=n+1;
       end
    end
end

size(entre)
size(comp)
xlswrite('entrenamiento',entre)
xlswrite('prueba',comp)
```

Anexo IV

Resultados de las pruebas de la mejor configuración de MVS

CORRIDAS DE CLASIFICACIÓN CON MVS DE LA MEJOR CONFIGURACIÓN
(n = 1500)

CORRIDA 1

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances         494               98.8    %
Incorrectly Classified Instances         6                1.2    %
Mean absolute error                      0.2512
Root mean squared error                  0.3117
Relative absolute error                697.1058 %
Root relative squared error            229.07   %
Total Number of Instances              500

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 480   1   0   0 |   a = clase1
   2   3   0   0 |   b = clase2
   2   0   1   1 |   c = clase3
   0   0   0  10 |   d = clase4
```

CORRIDA 2

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances         496               99.2    %
Incorrectly Classified Instances         4                0.8    %
Mean absolute error                      0.2513
Root mean squared error                  0.3136
Relative absolute error                960.7211 %
Root relative squared error            267.4731 %
Total Number of Instances              500

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 486   0   0   0 |   a = clase1
   2   2   0   0 |   b = clase2
   2   0   3   0 |   c = clase3
   0   0   0   5 |   d = clase4
```

CORRIDA 3

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances         494               98.8    %
Incorrectly Classified Instances         6                1.2    %
Mean absolute error                      0.251
Root mean squared error                  0.3127
Relative absolute error                937.3392 %
Root relative squared error            266.9068 %
Total Number of Instances              500

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 485   0   0   1 |   a = clase1
   4   1   0   0 |   b = clase2
   0   0   1   0 |   c = clase3
   1   0   0   7 |   d = clase4
```

CORRIDA 4

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances         993               99.3    %
Incorrectly Classified Instances         7                0.7    %
Mean absolute error                      0.2508
Root mean squared error                  0.3123
Relative absolute error               1077.9706 %
Root relative squared error            299.8454 %
Total Number of Instances             1000

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 975   1   1   1 |   a = clase1
   2   2   0   0 |   b = clase2
   1   0   8   0 |   c = clase3
   1   0   0   8 |   d = clase4
```

CORRIDA *5*

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances         987               98.7    %
Incorrectly Classified Instances        13                1.3    %
Mean absolute error                      0.2512
Root mean squared error                  0.3112
Relative absolute error                854.929  %
Root relative squared error            285.7903 %
Total Number of Instances             1000

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 972   1   0   3 |   a = clase1
   2   6   0   0 |   b = clase2
   7   0   2   0 |   c = clase3
   0   0   0   7 |   d = clase4
```

CORRIDA 6

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances         993               99.3    %
Incorrectly Classified Instances         7                0.7    %
Mean absolute error                      0.2508
Root mean squared error                  0.3123
Relative absolute error               1008.9415 %
Root relative squared error            287.3817 %
Total Number of Instances             1000

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 976   0   0   0 |   a = clase1
   3   2   0   0 |   b = clase2
   4   0   5   0 |   c = clase3
   0   0   0  10 |   d = clase4
```

CORRIDA 7

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        1980               99      %
Incorrectly Classified Instances        20                1      %
Mean absolute error                      0.2509
Root mean squared error                  0.3112
Relative absolute error                821.1962 %
Root relative squared error            272.5236 %
Total Number of Instances             2000

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 1942    0    1    4 |    a = clase1
    4    7    0    0 |    b = clase2
    5    0    5    2 |    c = clase3
    4    0    0   26 |    d = clase4
```

CORRIDA 8

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        1980               99      %
Incorrectly Classified Instances        20                1      %
Mean absolute error                      0.2512
Root mean squared error                  0.313
Relative absolute error               1016.9023 %
Root relative squared error            282.339  %
Total Number of Instances             2000

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 1945    1    2    2 |    a = clase1
    3    3    1    0 |    b = clase2
    9    0    9    0 |    c = clase3
    2    0    0   23 |    d = clase4
```

CORRIDA 9

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        1978               98.9    %
Incorrectly Classified Instances        22                1.1    %
Mean absolute error                      0.2512
Root mean squared error                  0.313
Relative absolute error                826.4294 %
Root relative squared error            237.7033 %
Total Number of Instances             2000

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 1927    0    1    1 |    a = clase1
    4    8    0    0 |    b = clase2
   10    0   15    0 |    c = clase3
    5    1    0   28 |    d = clase4
```

CORRIDA 10

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        3958               98.95   %
Incorrectly Classified Instances        42                1.05   %
Mean absolute error                      0.2509
Root mean squared error                  0.3111
Relative absolute error                850.9065 %
Root relative squared error            284.298  %
Total Number of Instances             4000

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 3891    3    2    7 |    a = clase1
   11   15    0    0 |    b = clase2
   13    0   12    2 |    c = clase3
    4    0    0   40 |    d = clase4
```

CORRIDA 11

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        3958               98.95   %
Incorrectly Classified Instances        42                1.05   %
Mean absolute error                      0.2513
Root mean squared error                  0.3132
Relative absolute error                906.2169 %
Root relative squared error            253.0767 %
Total Number of Instances             4000

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 3869    1    2    3 |    a = clase1
   15   15    0    0 |    b = clase2
   17    0   25    0 |    c = clase3
    4    0    0   49 |    d = clase4
```

CORRIDA 12

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        3961               99.025  %
Incorrectly Classified Instances        39                0.975  %
Mean absolute error                      0.251
Root mean squared error                  0.3127
Relative absolute error                958.5886 %
Root relative squared error            272.8387 %
Total Number of Instances             4000

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 3887    0    2    4 |    a = clase1
   13   11    0    1 |    b = clase2
   14    0   22    1 |    c = clase3
    4    0    0   41 |    d = clase4
```

CORRIDA 13

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        6829               98.885  %
Incorrectly Classified Instances        77                1.115  %
Mean absolute error                      0.251
Root mean squared error                  0.3113
Relative absolute error                836.6373 %
Root relative squared error            278.5737 %
Total Number of Instances             6906

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 6707    8    2   14 |    a = clase1
   18   23    0    0 |    b = clase2
   25    0   23    2 |    c = clase3
    8    0    0   76 |    d = clase4
```

CORRIDA 14

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        6828               98.8705 %
Incorrectly Classified Instances        78                1.1295 %
Mean absolute error                      0.2513
Root mean squared error                  0.3132
Relative absolute error                966.9261 %
Root relative squared error            268.7933 %
Total Number of Instances             6906

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 6695    4   11    5 |    a = clase1
   21   20    1    0 |    b = clase2
   26    0   36    0 |    c = clase3
    9    0    1   77 |    d = clase4
```

CORRIDA 15

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        6832               98.9285 %
Incorrectly Classified Instances        74                1.0715 %
Mean absolute error                      0.2511
Root mean squared error                  0.3128
Relative absolute error                945.6075 %
Root relative squared error            269.2117 %
Total Number of Instances             6906

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 6705    0    3    8 |    a = clase1
   28   17    1    1 |    b = clase2
   23    0   34    1 |    c = clase3
    8    1    0   76 |    d = clase4
```

CORRIDAS DE CLASIFICACIÓN CON MVS DE LA MEJOR CONFIGURACIÓN
(n = 198)

CORRIDA 1

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances         489               97.8    %
Incorrectly Classified Instances        11                2.2    %
Mean absolute error                      0.252
Root mean squared error                  0.3148
Relative absolute error                581.6052 %
Root relative squared error            282.5251 %
Total Number of Instances              500

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 482   0   2   4 |   a = clase1
   1   0   0   0 |   b = clase2
   2   0   1   0 |   c = clase3
   2   0   0   6 |   d = clase4
```

CORRIDA 2

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances         486               97.2    %
Incorrectly Classified Instances        14                2.8    %
Mean absolute error                      0.253
Root mean squared error                  0.3162
Relative absolute error                613.2039 %
Root relative squared error            259.0815 %
Total Number of Instances              500

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 482   2   1   0 |   a = clase1
   3   0   0   0 |   b = clase2
   5   0   0   1 |   c = clase3
   2   0   0   4 |   d = clase4
```

CORRIDA 3

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances         491               98.2    %
Incorrectly Classified Instances         9                1.8    %
Mean absolute error                      0.2527
Root mean squared error                  0.3161
Relative absolute error               1010.6667 %
Root relative squared error            279.6529 %
Total Number of Instances              500

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 487   0   0   0 |   a = clase1
   3   0   0   0 |   b = clase2
   3   0   0   0 |   c = clase3
   3   0   0   4 |   d = clase4
```

CORRIDA 4

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances         977               97.7    %
Incorrectly Classified Instances        23                2.3    %
Mean absolute error                      0.2529
Root mean squared error                  0.3163
Relative absolute error                565.9477 %
Root relative squared error            269.5012 %
Total Number of Instances             1000

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 962   2   5   4 |   a = clase1
   4   2   0   0 |   b = clase2
   6   0   3   0 |   c = clase3
   2   0   0  10 |   d = clase4
```

CORRIDA *5*

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances         989               98.9    %
Incorrectly Classified Instances        11                1.1    %
Mean absolute error                      0.2513
Root mean squared error                  0.3136
Relative absolute error                742.1743 %
Root relative squared error            362.3506 %
Total Number of Instances             1000

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 982   3   0   1 |   a = clase1
   4   1   0   0 |   b = clase2
   2   0   1   0 |   c = clase3
   1   0   0   5 |   d = clase4
```

CORRIDA *6*

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances         986               98.6    %
Incorrectly Classified Instances        14                1.4    %
Mean absolute error                      0.2522
Root mean squared error                  0.3153
Relative absolute error               1070.7827 %
Root relative squared error            296.2791 %
Total Number of Instances             1000

=== Confusion Matrix ===
   a   b   c   d   <-- classified as
 977   0   0   0 |   a = clase1
   5   0   0   0 |   b = clase2
   7   0   1   0 |   c = clase3
   2   0   0   8 |   d = clase4
```

CORRIDA 7

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        1951               97.55   %
Incorrectly Classified Instances        49                2.45   %
Mean absolute error                      0.253
Root mean squared error                  0.3166
Relative absolute error                584.2274 %
Root relative squared error            284.3097 %
Total Number of Instances             2000

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 1922    4   13   13 |    a = clase1
    5    9    0    0 |    b = clase2
   11    0    8    0 |    c = clase3
    3    0    0   12 |    d = clase4
```

CORRIDA 8

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        1964               98.2    %
Incorrectly Classified Instances        36                1.8    %
Mean absolute error                      0.2524
Root mean squared error                  0.3155
Relative absolute error                595.0186 %
Root relative squared error            248.9183 %
Total Number of Instances             2000

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 1929    3    2    1 |    a = clase1
   12    5    0    0 |    b = clase2
   12    2    7    0 |    c = clase3
    3    1    0   23 |    d = clase4
```

CORRIDA 9

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        1960               98      %
Incorrectly Classified Instances        40                2      %
Mean absolute error                      0.2532
Root mean squared error                  0.317
Relative absolute error               1012.7984 %
Root relative squared error            280.4136 %
Total Number of Instances             2000

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 1944    0    2    2 |    a = clase1
   16    0    0    0 |    b = clase2
   14    0    3    0 |    c = clase3
    6    0    0   13 |    d = clase4
```

CORRIDA 10

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        3909               97.725  %
Incorrectly Classified Instances        91                2.275  %
Mean absolute error                      0.2528
Root mean squared error                  0.3162
Relative absolute error                586.8557 %
Root relative squared error            286.6548 %
Total Number of Instances             4000

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 3857    5   22   22 |    a = clase1
   16    8    0    0 |    b = clase2
   18    0   10    1 |    c = clase3
    6    0    1   34 |    d = clase4
```

CORRIDA 11

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        3923               98.0803 %
Incorrectly Classified Instances        77                1.9197 %
Mean absolute error                      0.2523
Root mean squared error                  0.3154
Relative absolute error                635.261  %
Root relative squared error            272.8095 %
Total Number of Instances             4000

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 3876   17    3    2 |    a = clase1
   21    5    1    0 |    b = clase2
   17    1   12    1 |    c = clase3
   14    0    0   30 |    d = clase4
```

CORRIDA 12

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        3930               98.2548 %
Incorrectly Classified Instances        70                1.7452 %
Mean absolute error                      0.2525
Root mean squared error                  0.3158
Relative absolute error               1026.365  %
Root relative squared error            283.8887 %
Total Number of Instances             4000

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 3895    0    4    2 |    a = clase1
   21    0    0    0 |    b = clase2
   26    0    3    0 |    c = clase3
   17    0    0   32 |    d = clase4
```

CORRIDA 13

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        8020               97.6976 %
Incorrectly Classified Instances       189                2.3024 %
Mean absolute error                      0.2529
Root mean squared error                  0.3163
Relative absolute error                569.9514 %
Root relative squared error            272.6786 %
Total Number of Instances             8209

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 7895   11   46   41 |    a = clase1
   26   22    0    0 |    b = clase2
   45    0   20    1 |    c = clase3
   18    0    1   83 |    d = clase4
```

CORRIDA 14

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        8050               98.0631 %
Incorrectly Classified Instances       159                1.9369 %
Mean absolute error                      0.2524
Root mean squared error                  0.3153
Relative absolute error                636.2026 %
Root relative squared error            273.388  %
Total Number of Instances             8209

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 7951   27    5    8 |    a = clase1
   41    9    1    0 |    b = clase2
   43    2   19    1 |    c = clase3
   30    1    0   71 |    d = clase4
```

CORRIDA 15

```
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        8050               98.0631 %
Incorrectly Classified Instances       159                1.9369 %
Mean absolute error                      0.253
Root mean squared error                  0.3165
Relative absolute error                987.3217 %
Root relative squared error            273.5    %
Total Number of Instances             8209

=== Confusion Matrix ===
    a    b    c    d   <-- classified as
 7966    0    9   10 |    a = clase1
   52    0    0    1 |    b = clase2
   56    0   11    0 |    c = clase3
   31    0    0   73 |    d = clase4
```