IMPLEMENTACIÓN DE UNA SERIE DE PASOS PARA LA APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN EL ANÁLISIS DE INFORMACIÓN GENERADA POR LA PLANTA DEMEX DE ECOPETROL

TATIANA PÉREZ URIBE
OSWALDO JAVIER TARAZONA ROMÁN

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA
FACULTAD DE INGENIERÍA DE SISTEMAS
BUCARAMANGA

2006

IMPLEMENTACIÓN DE UNA SERIE DE PASOS PARA LA APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN EL ANÁLISIS DE INFORMACIÓN GENERADA POR LA PLANTA DEMEX DE ECOPETROL

TATIANA PÉREZ URIBE
OSWALDO JAVIER TARAZONA ROMÁN

Trabajo de investigación para optar el título de Ingeniero de Sistemas

Director
Ms.c Javier Hernández Cáceres

Asesor
Juan Carlos García Díaz
Ingeniero de Sistemas
Docente

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA
FACULTAD DE INGENIERÍA DE SISTEMAS
BUCARAMANGA

2006

Nota de aceptación:

_____

_____

_____

_____

_____

_____
Firma del Jurado

_____
Firma del Jurado

_____
Firma del Director

Bucaramanga, Noviembre 24 de 2006

AGRADECIMIENTOS

Queremos agradecer a la Universidad Autónoma de Bucaramanga por ofrecernos la oportunidad de hacer parte de los convenios de investigación con el Instituto Colombiano de Petróleo – ICP.

A la Ingeniera Martha Josefina Parra, codirectora del proyecto en el ICP, y al Ingeniero Juan Carlos García Díaz, nuestro asesor por su disponibilidad de tiempo, seguimiento y observaciones durante el transcurso de este proyecto.

A Javier Hernández Cáceres nuestro director por corregirnos y orientarnos en los momentos malos y buenos que surgieron durante el proceso de investigación, presentación y sustentación del proyecto.

A nuestros padres, hermanos y familiares por creer en nosotros, por sus sabios consejos y el continuo apoyo que día tras día hicieron posible este proyecto.

Y a todos los que de una u otra manera brindaron apoyo, colaboración, compañía, consejos para el desarrollo y terminación de este proceso de aprendizaje.

DEDICATORIAS

A Dios, a mis padres, a mis hermanos, a mi sobrina y familiares, por la paciencia, el amor y el apoyo que me han brindado a través de cada etapa de mi vida. A mis profesores por enseñarme lo que hoy sé y seré en un futuro no muy lejano. Por último y no con menor importancia, a mis amigos por su amistad, apoyo, compañía y cariño. Y a todos lo que de una u otra forma estuvieron conmigo en el paso por la Universidad.

TATIANA PÉREZ URIBE

A mi familia que me ha prestado apoyo incondicional, mis compañeros de carrera que me acompañaron en todos los momentos significativos, pero sobre todo a mi compañera Tatiana Pérez, que siempre sembró una luz de esperanza en los momentos más complicados de la realización de este proyecto. Finalmente agradezco a mi profesor y director de tesis Javier Hernández y al profesor Juan Carlos García, los cuales siempre me prestaron su colaboración incondicional.

OSWALDO JAVIER TARAZONA ROMÁN.

CONTENIDO

LISTA DE TABLAS

Pág.

LISTA DE FIGURAS

Pág.

LISTA DE ANEXOS

Pág.

GLOSARIO

MINERIA DE DATOS se define como un proceso analítico diseñado para explorar grandes volúmenes de datos denominados *datawarehouse*, con el objeto de descubrir patrones y modelos de comportamiento o relaciones entre diferentes variables. Esto permite generar un conocimiento que ayuda a mejorar la toma de decisiones en los procesos fundamentales de un negocio.

KDD (*Knowledge Discover and Data Mining*), "descubrimiento de conocimiento a partir de bases de datos", reconociendo patrones y asociaciones que se mantienen ocultas en los datos

DATAWAREHOUSE es una bodega donde están almacenados todos los datos necesarios para realizar las funciones de gestión de la empresa, de manera que puedan utilizarse fácilmente según se necesiten. El objetivo del *datawarehouse* es el de satisfacer los requerimientos de información interna de la empresa para una mejor gestión

JAVA lenguaje de objetos, independiente de la plataforma. Originalmente desarrollado por un grupo de ingenieros de Sun, su uso se destaca en el Web; sirve para crear todo tipo de aplicaciones (locales, Intranet o Internet).

WEKA (*Waikato Environment for knowledges Analysis*) es una colección de algoritmos para desarrollar tareas de Minería de Datos. *Weka* fue desarrollado en la Universidad de Waikato en Nueva Zelanda y es un *software open source* bajo GNU (*General Public License*).

GAMS es un poderoso paquete matemático que permite entre muchas opciones, el modelamiento de sistemas lineales, no lineales y mixtos, de programación entera, y problemas de optimización. Este paquete ha sido diseñado para trabajar problemas de gran magnitud, y para ser usado desde computadoras personales, hasta *mainframes* y supercomputadoras.

CRISP-DM provee una vista del ciclo de vida de los proyectos en Minería de Datos. Contiene las fases correspondientes a un proyecto sus respectivas tareas y las relaciones entre estas tareas. Estas relaciones pueden establecerse entre todas las tareas de la Minería de Datos y los objetivos, el campo de estudio y el interés del usuario dependiendo siempre de los datos.

RESUMEN

PALABRAS CLAVES: Minería de Datos, *WEKA*[1], GAMS[2], UML[3], RUP[4]

La información y los almacenes de datos como se conocen hoy en día, son probablemente uno de los recursos más valiosos para las empresas, ya que en ellos, reposa dormida información de carácter vital y altamente lucrativa, que además puede significar para las empresas reducciones de costos, y aumentos sustanciales en sus utilidades. La gran inquietud que mantiene a los expertos a la expectativa, es cómo sacar a la luz tales conocimientos y hacerlos efectivos para poner en marcha los planes futuros de las empresas. De aquí nació el concepto de KDD (*Knowledge Discovery and Data Mining*), o descubrimiento de conocimiento a partir de bases de datos, reconociendo patrones y asociaciones que se mantienen ocultas en los datos, esperando a ser extraídas para darles un uso adecuado.

Una de las compañías más interesadas en poder explotar sus datos recopilados a través de los años es el Instituto Colombiano de Petróleo ICP, el cual posee grandes almacenes de datos archivados en medios computacionales, esperando a ser explotados mediante las técnicas de Minería de Datos y técnicas estadísticas, las cuales generarán los conocimientos adecuados para las disminución de costos, y la maximización de la productividad.

En este proyecto se hace el máximo esfuerzo por mejorar un prototipo computacional ya existente, desarrollado en Java, y basado en el paquete computacional de Minería de Datos *WEKA*, con el cual se pretenden explorar los datos generados por la planta DEMEX de ECOPETROL, y que reposan en la base de datos SILAB. Se aplican los términos de Minería de Datos, y de ingeniería de software de la forma más fiel para obtener un producto de excelente calidad, y con resultados muy certeros, pero además para dar al usuario un soporte metodológico para aplicar correctamente las técnicas y así darle el mejor uso al prototipo que se presenta a continuación.

---

[1] *Waikato environment for knowledges analysis*
[2] *General algebraic modeling system*
[3] Lenguaje de modelado unificado
[4] Proceso Racional Unificado

INTRODUCCIÓN

Actualmente se cuenta con herramientas computacionales para la captura, pre-procesamiento de datos y aplicación de técnicas de Minería de Datos; estas herramientas computacionales se pueden encontrar libres y licenciadas, pero casi ninguna cumple satisfactoriamente con las expectativas de los clientes, ya que se basan en procesos exclusivamente asistidos por el usuario, y este simplemente quiere oprimir un botón y obtener los resultados. El software que se va a mejorar, el SPP (Sistema de Predicción de Propiedades), y en el proyecto no es una excepción a las herramientas actuales, es una herramienta más, pero con el gran valor agregado de que se entrega acompañado de un asistente metodológico que está guiando al usuario o investigador a través de los procesos, y proporciona consejos e información vital para poder seguir adelante con las investigaciones sobre Minería de Datos.

Como se puede apreciar en los objetivos, este proyecto está acompañado de un arduo trabajo de investigación en el cual intervienen expertos de diferentes disciplinas, que están a la expectativa de las mejoras y arreglos que se puedan agregar, es por eso que el ICP, por medio de un convenio con la UNAB, ha contado en manos de jóvenes investigadores, trabajos de este calibre, pensando en que las nuevas ideas y el espíritu de los estudiantes de esta institución saquen adelante y de forma eficiente estos trabajos.

El proyecto de grado consiste principalmente en realizar o diseñar una metodología que le indique al usuario cuál es el paso a seguir frente al prototipo computacional SPP así como la interpretación de cada una de las técnicas de Minería de Datos que en el prototipo se aplican, que son: ANOVA, Análisis de Regresión, Árboles de Decisión, Análisis de Componentes Principales, Análisis de Conglomerados o *Cluster*, Redes Bayesianas procesadas por el *WEKA* y el Análisis de Regresión no Lineal usando el GAMS.

A través de todo este trabajo de investigación se ha entrado al campo de la estadística, Minería de Datos y programación para comprender más a fondo el saber de nuestra profesión y ampliar aun más los campos de acción de ésta.

1. MARCO TEÓRICO

1.1 MINERÍA DE DATOS

El descubrimiento de conocimiento en base de datos (KDD) es la convergencia del Aprendizaje Automático, la Estadística, el Reconocimiento de Patrones, la Inteligencia Artificial, las Bases de Datos, la Visualización de Datos, los Sistemas para el Apoyo a la Toma de Decisiones, la Recuperación de Información, y otros muchos campos. KDD combina las técnicas tradicionales con numerosos recursos desarrollados en el área de la inteligencia artificial. En estas aplicaciones el término "Minería de Datos", está defina como un proceso analítico diseñado para explorar grandes volúmenes de datos (generalmente datos de negocio y mercado), denominados *datawarehouse*, con el objeto de descubrir patrones y modelos de comportamiento o relaciones entre diferentes variables.

Esto permite generar conocimiento que ayuda a mejorar la toma de decisiones en los procesos fundamentales de un negocio.[5] La Minería de Datos permite obtener valor a partir de la información que registran y manejan las empresas, lo que ayuda a dirigir esfuerzos de mejora respaldados en datos históricos de diversa índole.[6] Las consultas a esta gran bodega no son tan sistemáticas como las transacciones y usualmente demandan más recursos de cómputo. Resulta incluso conveniente separar los equipos y sistemas de la operación cotidiana de transacciones en línea del *datawarehouse*. La cual hace viable la revisión y el análisis de su información para el apoyo a las decisiones ejecutivas.[7]

Las herramientas OLAP[8] ofrecen un mayor poderío para revisar, graficar y visualizar información multidimensional, en características temporales, espaciales o propias.[9] Algunas posibilidades que ofrecen estas herramientas son:

- Mejorar el funcionamiento de la organización

---

[5] INFLEXA, ¿Qué es Minería de Datos? [Online, Articulo] 2003 santiago de chile. [Citado el 30 de enero 2006] Disponible en Internet: <http://www.inflexa.com/ jsp/template.jsp?pag=mineria-datos.htm&mnu=mnu-mineria.htm>

[6] ANSWERMATH, tutoriales de Minería de Datos [online, Tutorial] 2005. [Citado el 8 de Febrero]. Disponible en Internet: <http://www.answermath.com/ mineria_de_datos.htm>

[7] DAEDALUS - DATA, Decisions and Language, S. A. Minería de Datos [online, Articulo] 2006. [Citado el 24 de febrero 2066]. Disponible en Internet: <http://www.daedalus.es/AreasMD-E.php> y <http://www.daedalus.es/AreasMD Fases-E.php>

[8] OLAP: On-Line Analytical Processing

[9] Ibid., 16

- Optimizar el manejo de sus bases de datos
- Predicción automatizada de tendencias y comportamientos
- Obtener ventajas comerciales
- Mejorar calidad de productos
- Descubrimiento automatizado de modelos desconocidos
- Descubrimiento de anomalías y acciones fraudulentas por parte de clientes

La Minería de Datos puede ser dividida en:[10]

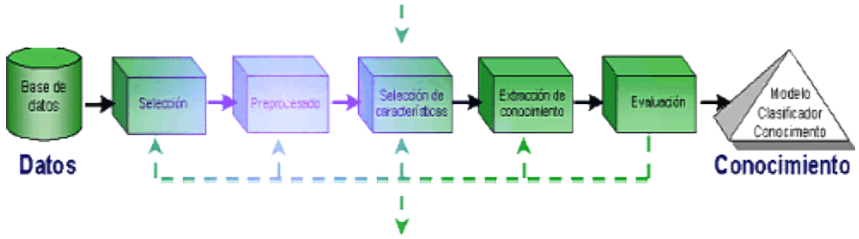
Minería de Datos **predictiva (MDP):** usa primordialmente técnicas estadísticas.

Minería de Datos **para descubrimiento de conocimiento (MDDC):** usa principalmente técnicas de inteligencia artificial.

1.1.1 Fases de la Minería de Datos

- Filtrado de datos

Figura 1. Filtrado de datos



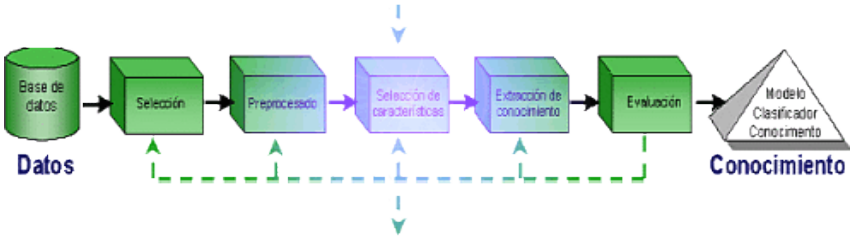
Fuente: DAEDALUS - DATA, Decisions and Language, S. A. Minería de Datos [online, Articulo] 2006. [Citado el 24 de febrero 2066]. Disponible en Internet: <http://www.daedalus.es/AreasMD-E.php> y <http://www.daedalus.es/AreasMD Fases-E.php>

[10] BRESSÁN, Griselda. Trabajo monográfico de adscripción. Lic. En Sistemas de Información Almacenes de Datos y Minería de Datos. [Online, Articulo], 2003. [Citado el 5 de febrero de 2006] Disponible en Internet: <http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos /MineriaDatos Bressan.htm>

Mediante el preprocesado se eliminan valores incorrectos, no válidos y desconocidos según las necesidades y el algoritmo a usar, se obtienen muestras de los mismos o se reduce el número de valores posibles.

– Selección de variables

Figura 2. Selección de Variables



Fuente: Ibíd., p. 17

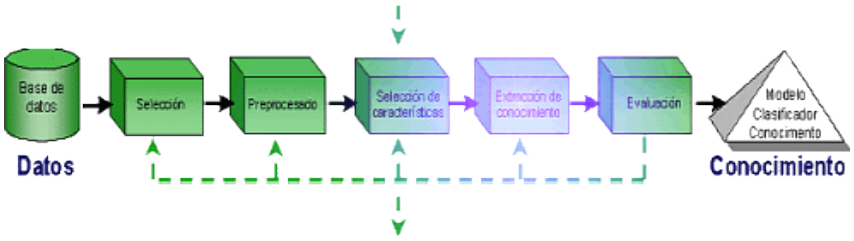
La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema.

Los métodos para la selección de características son básicamente dos:

Aquellos basados en la elección de los mejores atributos del problema y aquellos que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos.

– Algoritmo de extracción de conocimientos

Figura 3. Algoritmos de Extracción de Conocimientos

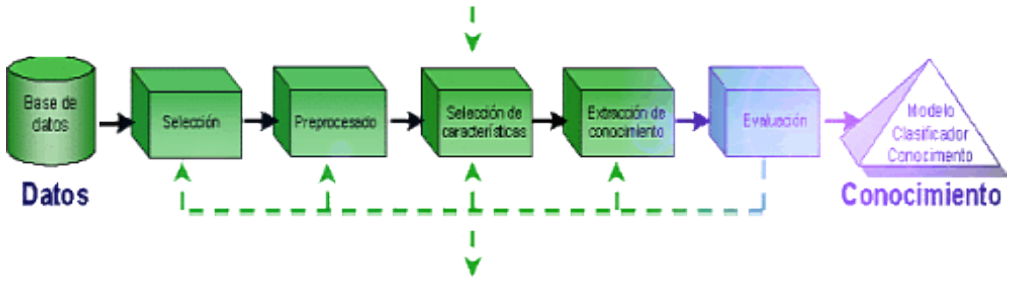


Fuente: Ibíd., p. 17

Mediante la aplicación de una técnica de Minería de Datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.

– Interpretación y evaluación

Figura 4. Interpretación y evaluación



Fuente: Ibíd., p. 17

Una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Las técnicas de Minería de Datos se emplean para mejorar el rendimiento de procesos de negocio o industriales en los que se manejan grandes volúmenes de información estructurada y almacenada en bases de datos. Asimismo, la Minería de Datos es fundamental en la investigación científica y técnica, como herramienta de análisis y descubrimiento de conocimiento a partir de datos de observación o de resultados de experimentos. Para realizar un proyecto en Minería de Datos se deben seguir unas fases, las cuales son siempre las mismas, independientemente de la técnica específica de extracción de conocimiento usada.[11]

[11] DAEDALUS - DATA, Decisions and Language, S. A. Minería de Datos [online, Articulo] 2006. [Citado el 24 de febrero 2066]. Disponible en Internet: <http://www.daedalus.es/AreasMD-E.php> y <http://www.daedalus.es/AreasMD Fases-E.php>

Las prácticas de Minería de Datos se realizan con base en procedimientos como:[12]

- **Clasificación**. Consiste en examinar las características de una entidad nueva y asignarle una clase predefinida. Por ejemplo: clasificar a un nuevo cliente según su riesgo de crédito (alto, medio, bajo).

- **Estimación**. Consiste en examinar las características de una entidad nueva y asignarle una clase predefinida. Por ejemplo: ingresos, balance de tarjetas de crédito, etc.

- **Predicción**. Predicción de fidelidad de clientes.

- ***Clustering***. Tiene como objetivo el segmentar a un grupo diverso en un conjunto de subgrupos o "cluster". A diferencia de clasificación, *clustering* no depende de clases predefinidas. Y es el primer paso en segmentación de mercado. Por ejemplo: un cluster particular de síntomas puede indicar una enfermedad particular.

- **Descripción y visualización**. Algunas veces el objetivo es simplemente describir qué esta ocurriendo en una base de datos compleja, para así aumentar el entendimiento de las personas, productos o procesos que generaron los datos inicialmente.

1.1.2 Aplicaciones existen una gran cantidad de aplicaciones,[13] en áreas tales como:

- **Astronomía**: clasificación de cuerpos celestes

- **Aspectos climatológicos**: predicción de tormentas, etc.

- **Medicina**: caracterización y predicción de enfermedades, probabilidad de respuesta satisfactoria a tratamiento médico

- **Industria y manufactura**: diagnóstico de fallas

- **Mercadotécnia**: identificación de clientes susceptibles de responder a ofertas de productos y servicios por correo, fidelidad de clientes, selección de sitios de tiendas, afinidad de productos, etc.

[12] INFLEXA, Op. Cit 16
[13] BRESSÁN, Griselda Op. Cit 17

- **Inversión en casas de bolsa y banca**: análisis de clientes, aprobación de préstamos, determinación de montos de crédito, etc.

- **Detección de fraudes y comportamientos inusuales**: telefónicos, seguros, en tarjetas de crédito, de evasión fiscal, electricidad, etc.

- **Segmentación de mercado** (*clustering*)

- Determinación de niveles de audiencia de programas televisivos

- Normalización automática de bases de datos

1.1.3 ¿Qué promete la Minería de Datos? Que exista una reacción del público por el uso indiscriminado de datos personales para ejercicios de Minería de Datos. También es muy posible que se desee hacer inferencias y análisis de datos sobre un periodo determinado, pero que durante dicho periodo no se haya registrado el mismo número de variables, o que éstas no tengan la misma precisión, o carezcan de la misma interpretación.

- Resulta un buen punto de encuentro entre los investigadores y las personas de negocios.

- Ahorra grandes cantidades de dinero a una empresa y abre nuevas oportunidades de negocios.

- Trabajar con esta tecnología implica cuidar un sinnúmero de detalles, debido a que el producto final involucra "toma de decisiones".

- Contribuye a la toma de decisiones tácticas y estratégicas proporcionando un sentido automatizado para identificar información clave desde volúmenes de datos generados por procesos tradicionales y de *e-Business.*

- Permite a los usuarios dar prioridad a decisiones y acciones.

- Genera modelos descriptivos: en un contexto de objetivos definidos en los negocios permite a las empresas explorar automáticamente, visualizar y comprender los datos e identificar patrones, relaciones y dependencias que impactan en los resultados finales.

- Genera modelos predictivos: permite que relaciones no descubiertas e identificadas a través del proceso de Minería de Datos sean expresadas como reglas de negocio.

1.2 ALGORITMOS DE MINERÍA DE DATOS

Se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento. Los algoritmos supervisados o predictivos predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos (atributos descriptivos). A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos.

Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como aprendizaje supervisado y se desarrolla en dos fases: entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos). Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución predictiva, en ese caso hay que recurrir a los métodos no supervisados o del descubrimiento del conocimiento que descubren patrones y tendencias en los datos actuales. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio de ellas.

Tabla 1. Clasificación de las técnicas de Minería de Datos

Supervisados	**No supervisados**
Árboles de decisión	Detección de desviaciones
Inducción neuronal	Segmentación
Regresión	Agrupamiento ("clustering")
Series temporales	Reglas de asociación
	Patrones secuenciales

Fuente: MORENO GARCÍA, María. MIGUEL QUINTALES, Luís. GARCÍA PEÑALVO, Francisco y POLO MARTÍN, José. Aplicación de técnicas de Minería de Datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software [online, Articulo]. [Citado el 11 de febrero 2006]. Disponible en Internet: <http://www.sc.ehu.es/jiwdocoj/ remis/docs/minerw.pdf> P.3

1.2.1 Análisis de conglomerados o *clustering* el Análisis de *Clusters* se utiliza para clasificar los objetos o casos en grupos relativamente homogéneos llamados conglomerados (*clusters*). Los objetos en cada conglomerado tienden a ser similares entre sí y diferentes a los objetos de los otros grupos con respecto a algún criterio de selección predeterminado. De este modo, si la clasificación es un éxito, los objetos dentro del cluster estarán muy cercanos unos de otros en la representación geométrica, y los clusters diferentes estarán muy apartados.

Este análisis no hace ninguna distinción entre variables dependientes (VD[14]) y variables independientes (VI[15]) sino que calcula las relaciones interdependientes de todo el conjunto de variables.

El Análisis de Clusters presenta un fuerte contraste con el análisis de la varianza, la regresión, el análisis discriminante y el análisis factorial, que se basan en un razonamiento estadístico.

Principios fundamentales

- Informe de Aglomeración: Ofrece información sobre los objetos o casos que se combinan en cada etapa de un proceso de agrupación jerárquica.

- Centroides de Agrupamiento: Son los valores medios de las variables para todos los casos u objetos de un grupo particular.

- Centros de Agrupamiento: Son los puntos de partida iniciales en la agrupación no jerárquica. Los grupos se construyen alrededor de estos centros o semillas.

- Participación en el Grupo: Indica el grupo al que pertenece cada objeto o caso.

- Pasos del análisis de conglomerados

Formulación del Problema: el conjunto de variables seleccionado debe describir la similitud entre los objetos en términos relevantes para el problema de investigación.

Selección de una medida de similitud: La similitud es una medida de correspondencia o semejanza entre los objetos que van a ser agrupados. La estrategia más común consiste en medir la equivalencia en términos de la distancia entre los pares de objetos.

En la medición de la similitud entre los objetos de un Análisis de *Clusters* existen tres métodos: Medidas de Correlación, Medidas de Distancia y Medidas de Asociación

Las medidas de correlación y las de distancia requieren datos métricos, mientras que las medidas de asociación requieren datos no métricos.

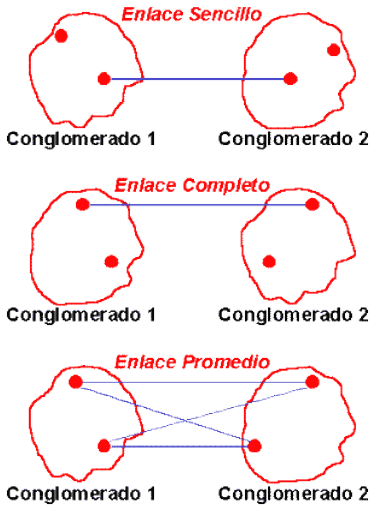
[14] VD: Variable Dependiente
[15] VI: Variable Independiente

Selección del procedimiento de agrupación: hay dos tipos de procedimientos: jerárquicos y no jerárquicos.

Los métodos jerárquicos pueden ser por aglomeración o por división. Los conglomerados se forman al agrupar los objetos en conjuntos cada vez más grandes. Este proceso continúa hasta que todos los objetos formen parte de un solo grupo. El conglomerado por división comienza con todos los objetos agrupados en un solo conjunto. Los conglomerados se dividen hasta que cada objeto sea un grupo independiente.

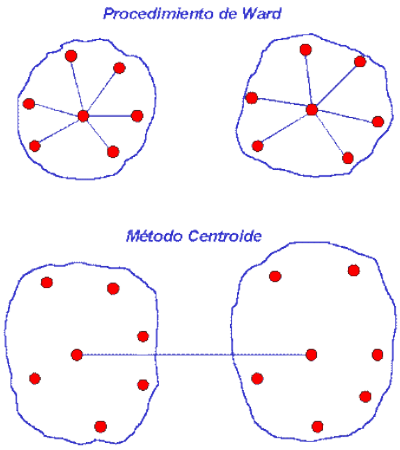
El método del enlace completo es similar al enlace sencillo, la distancia entre dos conglomerados se calcula como la distancia entre sus puntos más lejanos. En el método del enlace promedio la distancia entre dos conglomerados se define como el promedio de las distancias entre todos los pares de objetos, donde se encuentra un miembro del par de cada uno de los conglomerados (Ver Figura 5 Métodos de Enlace para el Conglomerado).

Figura 5. Método de enlace para el conglomerado



Fuente: Estadístico, es el sitio Web especialistas en consultoría y formación estadística, integrado por expertos en los programas SPSS, SAS, CLEMENTINE entre otros, en la Web de data Mining Institute encontrará todo lo referente a la estadística: cursos, artículos, software, enlaces, consultoría, libros, diccionario estadístico y tests. [Online, Artículo estadístico] 2004. [Citado el 07 de Febrero 2006] Disponible en <http://www.estadistico.com/arts.html?20001023>

Figura 6. Otros métodos de Agrupación por Aglomeración



Fuente: Ibíd., p. 24

En el Método Centroide, la distancia entre dos grupos es la distancia entre sus centroides, como se muestra en la Figura 6, otros métodos de Agrupación por Aglomeración. Cada vez que se agrupan los objetos, se calcula un centroide nuevo.

Los métodos de conglomerados no jerárquicos, con frecuencia se conocen como Agrupación de K Medias. Estos métodos incluyen el Umbral Secuencial, Umbral Paralelo y la División para la Optimización. En el método del Umbral Secuencial, se selecciona un centro de grupo y se agrupan todos los objetos dentro de un valor de umbral que se especifica previamente a partir del centro. El método del Umbral Paralelo funciona de manera similar, excepto que se seleccionan simultáneamente varios centros de grupo y se agrupan los objetos del nivel del umbral dentro del centro más próximo. El método de División para la Optimización difiere de los otros dos procedimientos de umbral en que los objetos pueden reasignarse posteriormente a otros grupos, a fin de optimizar un criterio general, como la distancia promedio dentro de los grupos para un número determinado de conglomerados.

Decisión del Número de Conglomerados. Un gran problema en todas las técnicas de aglomeración es cómo seleccionar el número de grupos (*clusters*). Para el caso del análisis *cluster* jerárquico, las distancias existentes entre los *clusters* reflejadas en las distintas etapas del proceso de aglomeración puede servir de guía útil, el analista podría así establecer un tope para detener el proceso a su conveniencia.

Para el caso del análisis *cluster* no jerárquico, se puede trazar un gráfico que compare el número de grupos con la relación entre la varianza total de los grupos y la varianza entre los grupos. Los investigadores deben examinar la variación producida entre los tamaños de los grupos desde una perspectiva conceptual, comparando los resultados obtenidos con las expectativas creadas en los objetivos del estudio.

1.2.2 Análisis factorial el análisis factorial y el análisis de componentes principales están muy relacionados; algunos autores consideran el segundo como una etapa del primero y otros los consideran técnicas diferentes. En el proyecto de investigación que se está realizando se optó por la segunda opción, es decir, considerar el análisis de componentes principales como una etapa del análisis factorial.

El análisis factorial busca factores que expliquen la mayor parte de la varianza común y se diferencia la varianza común, y la varianza única; la varianza común es parte de la variación de la variable que es compartida con las otras variables, y la varianza única es parte de la variación de la variable que es propia de esa variable. El análisis factorial pretende hallar un nuevo conjunto de variables, menor en número que las variables originales, que exprese lo que es común a esas variables. También se puede decir que el análisis factorial supone que existe un factor común subyacente a todas las variables, y en el análisis de componentes principales no se tiene en cuenta esto.

Las técnicas modernas para el desarrollo de un análisis factorial se basan en una estrategia de cuatro pasos, los cuales se mencionaran a continuación:

Primer Paso: consta del establecimiento de los objetivos del análisis factorial, identificación de estructura mediante el resumen de datos, la propia reducción de los datos, uso del análisis factorial con otras técnicas, y la selección de las nuevas variables.

Segundo Paso: análisis de correlaciones entre las variables y los encuestados, medición y selección de variables, y análisis del tamaño muestral.

Tercer Paso: se da paso al análisis de los resultados en la matriz de correlaciones anti-imagen, contraste de esfericidad de Bartlett.

Cuarto paso: criterios para la significaron de las cargas factoriales, interpretación de la matriz de factores.

1.2.2.1 Análisis de componentes principales trata de hallar componentes (factores) que sucesivamente expliquen la mayor parte de la varianza total. El

análisis de componentes principales no hace distinción entre los dos tipos de varianza (común y única), se centra en la varianza total. El análisis de componentes principales trata de hallar combinaciones lineales de las variables originales que expliquen la mayor parte de la variación total. Cuando se realiza el análisis de componentes principales, la primera componente (o factor) es aquel que explica una mayor parte de la varianza total, el segundo factor explica la mayor parte de la varianza total restante, y así sucesivamente; por consiguiente sería posible obtener tantos componentes como variables originales, pero esto en la práctica no tendría ningún sentido, se recomiendan tres factores.

El análisis de componentes principales está definido como "modelo factorial en el que los factores se basan en la varianza total". En el análisis de componentes principales, se usan las unidades que aparecen en la diagonal de la matriz de correlación; este procedimiento implica, por lo que se refiere al cálculo, que toda la varianza es común o compartida.[16] Una matriz de correlaciones está definida como "Tabla que indica las intercorrelaciones entre todas las variables"[17]

Este método de análisis permite la estructuración de un conjunto de datos multivariados obtenidos de una población, cuya distribución de probabilidades no necesita ser conocida. Se trata de una técnica matemática que no requiere un modelo estadístico para explicar la estructura probabilística de los errores. Se puede suponer que la muestra tiene una distribución multinormal, de esta manera se estudia la significación estadística y es posible utilizar la muestra efectivamente observada, para efectuar pruebas de hipótesis que contribuyan a conocer la estructura de la población original con un cierto grado de confiabilidad, fijado a priori o a posteriori.

Objetivos: los objetivos más importantes del análisis de componentes principales son:[18]

- Generar nuevas variables que puedan expresar la información contenida en el conjunto de datos original.

- Reducir la dimensión del problema que se esta estudiando, como paso previo para futuros análisis.

- Eliminar algunas de las variables originales si ellas aportan poca información.

---

[16] HAIR, ANDERSON; TATHAM y BLACK. Análisis Multivariante, quinta edición, Prentice Hall, 2001, p. 143-148, 347-349, 767, 779

[17] HAIR, Op. Cit. p. 779

[18] PLA, Laura. Análisis de multivariado: método de componentes principales. Washington: Secretaría General de la Organización de los Estados Americanos, Programa Regional de Desarrollo Científico y Tecnológico. p. 1-17.1986. Serie de Matemáticas, monografía No 27.

El análisis de componentes principales ha sido aplicado en psicología, medicina, meteorología, geografía, ecología y agronomía. Se usa cuando se desea conocer la relación entre los elementos de una población y se sospecha que en dicha relación influye de manera desconocida un conjunto de variables o propiedades de los elementos.

Al estudiar un conjunto de *n* individuos mediante *p* variables es posible encontrar nuevas variables denominadas y (*k*), *k*= 1,…, p que sean combinaciones lineales de las variables originales *x* (*j*). Esto implica encontrar (*p* * *p*) constantes tales que son *l* (*jk*) en la Ecuación 1.

$$y(k) = \sum_{j=1}^{p} l(jk)\, x(j), \qquad k = 1,..., p \qquad \textbf{(1)}$$

1.2.3 Análisis de regresión el análisis de regresión múltiple (con dos o más variables independientes) es una técnica estadística general utilizada para analizar las relaciones entre una única variable criterio y varias variables independientes; su formulación básica está expresada en la Ecuación 2:

$$Y_1 = X_1 + X_2 + ... + X_n \qquad \textbf{(2)}$$

El análisis de regresión múltiple puede utilizarse para estudiar la relación entre una única variable criterio (criterio) y varias variables independientes (predictores). El objetivo es usar las variables independientes cuyos valores son conocidos, para predecir la única variable criterio seleccionada por el investigador.[19]

Cada variable predictor es ponderada, de forma que las ponderaciones indican su contribución relativa a la predicción conjunta. Al calcular las ponderaciones, el procedimiento del análisis de regresión asegura la máxima predicción a partir de un conjunto de variables independientes. Estas ponderaciones facilitan también la interpretación de la influencia de cada variable en la realización de la predicción.

El conjunto de variables independientes ponderadas es conocido también como valor teórico de la regresión, una combinación lineal de las variables independientes que predice mejor la variable dependiente. La Ecuación de

[19] HAIR, Op. Cit p. 143-148

regresión, también denominada valor teórico de la regresión, es el ejemplo de valor teórico más ampliamente reconocido entre todas las técnicas multivariantes.

En el análisis de regresión múltiple se tiene en cuenta el impacto de la multicolinealidad, lo que se refiere a la correlación entre tres o más variables independientes (evidenciada cuando se hace la regresión de una respecto de las otras). El impacto de la multicolinealidad consiste en reducir el poder predictivo de cualquier variable independiente individual en la medida en que se está asociando con las otras variables independientes. Conforme aumenta la colinealidad, la varianza única explicada por cada variable independiente se reduce y el porcentaje de predicción compartida aumenta.

1.2.4 Árboles de decisión el conocimiento obtenido en el proceso de aprendizaje se representa mediante un árbol en el cual cada nodo interior contiene una pregunta sobre un atributo concreto (un hijo por cada posible respuesta) y cada hoja del árbol se refiere a una decisión. Un árbol de decisión puede usarse para clasificar un caso comenzando desde su raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos hasta que encuentra una hoja del árbol.

La construcción de los árboles de decisión se hace recursivamente de forma descendente, por lo que se emplea el acrónimo TDIDT (top-*down induction on decision trees*) para referirse a la familia completa de algoritmos de este tipo. La familia de algoritmos TDIDT abarca desde algoritmos (ver Figura 7) como CLS, ID3, C4.5 o CART (*Classification and regression trees*). Los algoritmos TDIDT suelen presuponer que no existe ruido en los datos de entrada e intentan alcanzar una descripción perfecta de los mismos.

Una vez construido el árbol de decisión completo que se adapta perfectamente al conjunto de datos, la representación del conocimiento mediante árboles de decisión es bastante simple, a pesar de carecer de la expresividad de las redes semánticas o de la lógica de primer orden, se utiliza muy a menudo para resolver problemas de clasificación de todo tipo.[20]

Problemas en los que se aplican árboles de decisión:

- Las instancias se representan como parejas atributo-valor
- La función objetivo toma valores discretos

[20] BERZAL GALIANO, Fernando y TALAVERA CUBERO. Departamento de ciencias de la computación e inteligencia artificial, ETS-Ingeniería informática, Universidad de Granada [online, Articulo]. [Citado el 20 de febrero 2006]. Disponible en Internet: <http://elvex.ugr.es/etexts/spanish/proyecto/cap5.pdf> P. 3.

- Las descripciones requieren disyunciones
- El conjunto de entrenamiento puede contener errores
- Las instancias de entrenamiento pueden no tener todos los atributos

Figura 7. Algoritmos Árboles de Decisión

Fuente: Departamento de Sistemas Informáticos y Computación. Valencia España. Aprendizaje de árboles de decisión [online, Articulo]. [Citado el 22 de febrero 2006]. Disponible en Internet: <http://www.dsic.upv.es/asignaturas/facultad/apr/ decision.pdf> P. 5.

1.2.4.1 Algoritmo ID3

- Construye el árbol de manera *top-down*[21](de arriba hacia abajo)
- En cada paso se pregunta:
- ¿Cuál atributo se debe probar en cada nodo? Por cada valor del atributo seleccionado, se genera una rama y se repite el proceso en cada una de ellas tomando sólo las instancias que tienen el valor del atributo correspondiente a la rama. Nunca se verifica que el atributo seleccionado haya sido realmente el mejor.

[21] Instituto de Computación, Universidad de la República, Montevideo-Uruguay. Árboles de decisión [online, Articulo]. [Citado el 22 de febrero 2006]. Disponible en Internet: < http://www.fing.edu.uy/inco/cursos/aprendaut/transp/arboles.pdf> P.8

Crear una raíz. Si todos los ejemplos. Son positivos → etiquetar con +
Si todos los ejemplos son negativos → etiquetar con -
Si no queda atributos → etiquetar con el valor más común

En caso contrario:

A = atributo que mejor clasifica los ejemplos
Hacer que la raíz pregunte por A
Para cada valor v_i de A, se genera una rama

1.2.4.2 Algoritmo C4.5. El C4.5 es una extensión del ID3, y es el mismo J48 de *WEKA* que permite trabajar con valores contínuos para los atributos, separando los posibles resultados en dos ramas: una para aquellos $A_i<=N$ y otra para $A_i>N$. Este algoritmo fue propuesto por Quinlan en 1993. El algoritmo C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente. El árbol se construye mediante la estrategia de profundidad-primero *(depth-first).* El algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información. Para cada atributo discreto, se considera una prueba con *n* resultados, siendo *n* el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una *prueba binaria* sobre cada uno de los valores que toma el atributo en los datos.[22]

1.2.4.3 Entropía y ganancia de información es una manera de cuantificar la bondad de un atributo en este contexto, consiste en considerar la cantidad de información que proveerá este atributo. En general, si los posibles valores del atributo v_i ocurren con probabilidades $P(v_i)$, entonces el contenido de información o entropía *E* de la respuesta actual, está dado por[23] la Ecuación 3:

$$E(P(v_i),\ldots,P(v_n)) = \sum_{i=1}^{n} -P(v_i)\ \log_2\ P(v_i) \qquad \textbf{(3)}$$

[22] SERVENTE, Magdalena. Algoritmos TDIDT aplicados a la Minería de Datos inteligente. 2002 [online, Articulo]. [Citado el 21 de octubre 2006]. Disponible en Internet: <http://www.fi.uba.ar/laboratorios/lsi/servente-tesisingenieriainformatica.pdf > p. 77-89.

[23] HERNANDEZ GUERRA, Alejandro. Aprendizaje Automático: Árboles de Decisión. Universidad Veracruzana, México. 2004. [online, Articulo]. [Citado el 24 de febrero 2006]. Disponible en Internet: <http://www.uv.mx/aguerra/teaching/ MIA/MachineLearning/clase07.pdf> p. 6-8.

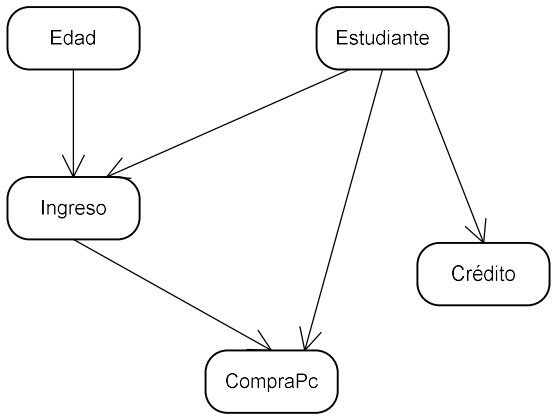
Los árboles de decisión son adecuados cuando

- Las instancias del concepto son representadas por pares atributo-valor
- La función objetivo tiene valores de salida discretos
- Las descripciones del objeto son disjuntivas
- El conjunto de aprendizaje tiene errores
- El conjunto de aprendizaje es incompleto

1.2.5 Redes bayesianas las redes bayesianas nacieron como una aportación de diferentes campos de investigación: teoría de toma de decisiones, estadística e inteligencia artificial. Las redes bayesianas representan el conocimiento cualitativo del modelo mediante un grafo dirigido acíclico, es decir, no contiene ningún ciclo simple. Además expresan de manera numérica las relaciones entre las variables. Esta parte cuantitativa suele especificarse mediante distribuciones de probabilidad como una medida de la carencia que se tiene sobre las relaciones entre variables del modelo.

Formalmente, una red bayesiana es una tupla (una lista con un número limitado de objetos) $B = (G,\Theta)$, donde G es el grafo y Θ es el conjunto de la distribución de probabilidad $P(X_i \mid Pa(X_i))$ para cada variable desde i = 1 hasta n y $Pa(X_i)$ representa los padres de la variable X_i en el grafo G. Un ejemplo de una red bayesiana puede verse en la Figura 8, donde se observa una red que refleja las relaciones entre las variables de un pequeño dominio para determinar si un cliente comprará o no una computadora personal.

Figura 8. Red bayesiana



Fuente: HERNÁNDEZ ORALLO, José; RAMÍREZ QUINTANA; Maria José y FERRI RAMÍREZ, Cesar. Introducción a la Minería de Datos. 2005. Editorial Pearson, p. 266-269

En la Figura 8 se observa que el nivel de ingresos depende de la edad y de su condición de estudiante, la compra de un PC personal depende de la edad, condición del estudiante y el nivel de crédito. Una relación indirecta es la edad ya que es independiente del nivel de crédito, y que los ingresos son independientes de comprar una computadora personal conocida la edad. Para representar la parte cuantitativa del modelo de la red se cita la teoría de la probabilidad. Una distribución de probabilidad P puede considerarse como la relación representada en la Ecuación 4:

$$I(X,Y|Z) \Leftrightarrow P(X|YZ) = P(X|Z) \quad \textbf{(4)}$$

Donde X, Y, Z son subconjuntos de variables del modelo y la sentencia $I(X,Y|Z)$ se interpreta como una relación de independencia condicional. Se leería: X es condicionalmente independiente de Y conocido Z. Una red bayesiana hace que la distribución de probabilidad conjunta se pueda almacenar de una manera mucho más eficiente y local a cada una de las variables de la siguiente manera ver Ecuación 5.

$$P(X1,...,Xn) = \prod_{i=1}^{n} P(X_i|Pa(X_i)) \quad \textbf{(5)}$$

1.2.5.1 Aprendizaje de redes bayesianas el problema del aprendizaje de redes bayesianas se puede definir como: “dado un conjunto de datos D encontrar el grafo dirigido acíclico G que mejor represente el conjunto de dependencias/independencias presentes en los datos”. Para solucionar este problema, es decir, calcular la probabilidad *a posteriori* (Ecuación 6) de una red bayesiana en concreto, dado el conjunto de datos conocidos, esto es, P (G|D). Una vez calculada se puede comparar en cada grafo y escoger la mejor.

$$P(G \mid D) = \frac{P(D \mid G)P(G)}{P(D)} \quad \textbf{(6)}$$

Algoritmo de Búsqueda K2 este algoritmo parte del conjunto de padres para cada variables es el conjunto vacío. Posteriormente, pasa a procesar cada variable X_i, calcula la ganancia que se produce en la medida utilizada al introducir una variable X_j como padre de X_i, esto es, para todo j<i y que se queda con la que produce la mejor ganancia. Desde el enfoque de la red bayesiana, esta ganancia se calcula como la razón de la métrica vista en el punto anterior de una red bayesiana con la variable X_i sin padres y una red bayesiana donde X_i tenga como padre X_j.

El pseudocógido puede ser visto en la Figura 9.

Figura 9. Pseudocódigo del algoritmo k2

```
ALGORITMO  K2(X:nodos (variables ordenadas), D:Datos)
  Fase de inicialización:
   Para cada Xi (i=0 hasta n)
     Pa(Xi) 0 conjunto vacío
   Fin para cada
  Fase iterativa:
   Para cada Xi (i=0 hasta n)
    ok := true
    Mientras ok
         Sea Xj el nodo tal que j<i y Xj no pertenezca a Pa(Xi) que maximiza
                                   fi(Xi | Pa(Xi) U Xj:D).
         Si fi(Xi | Pa(Xi) U Xj:D) > fi(Xi | Pa(Xi):D) entonces Pa(Xi):=Pa(Xi)U Xj
              En caso contrario ok:=false
    Fin mientras
  Fin para cada
Fin algoritmo
```

Fuente: HERNÁNDEZ, Op. Clt p. 267-268

1.3 ANOVA

El nivel de error tipo I (denotado por α o nivel de significación), que indica el nivel de probabilidad del investigador aceptara para concluir que las medias de los grupos son diferentes cuando realmente no lo son. ANOVA evita el aumento del error tipo I al comparar un conjunto de grupos de tratamiento, determinado si el conjunto completo de medidas muéstrales indica que las muestras fueron tomadas de la misma población general. Es decir, ANOVA es empleado para determinar la probabilidad de que las diferencias en las medidas entre varios grupos sean debidas solamente al error muestral. La lógica de un contraste ANOVA compara dos cálculos independientes de la varianza para la variable independiente uno que refleja la variabilidad general de los encuestados entre los grupos (CMI) y otro que representa las diferencias entre los grupos que se atribuyen a los efectos del tratamiento (CME).

Cálculo de la varianza entre los grupos (CMI: cuadro medio intra grupos): es una estimación de la variabilidad aleatoria de los encuestados sobre las variables dependiente dentro de un grupo de tratamiento, se basa en desviaciones de puntuaciones individuales respecto de las medias de sus grupos respectivas. Se le denomina también varianza del error.

Cálculo de la varianza entre grupos (CME: cuadrado medio entre grupos): se basa en desviaciones de las medias de los grupos de tratamiento sobre la media global de todas las puntuaciones. Bajo la hipótesis nula de que no hay efectos de tratamiento esta varianza refleja cualquier efecto de tratamiento que exista.

Dado que la hipótesis nula de no existencia de diferencias entre los grupos es cierta, los cuadrados medios intra y entre grupos representan estimaciones independientes de la varianza poblacional. Su ratio es una medida de cuánta varianza es atribuible a los diferentes tratamientos frente a la varianza esperada del muestreo aleatorio. Este ratio proporciona un valor de un estadístico F, cuyo cálculo se parece al de valor t dada la Ecuación 7:

$$EstadísticoF = \frac{CM_E}{CM_I} \quad \textbf{(7)}$$

Valores grandes del estadístico F (Fcrit) llevan al rechazo de la hipótesis nula de que no existen diferencias en las medias de los grupos. Para determinar este rechazo se halla el valor crítico para el estadístico F atendiendo a la distribución F con (k-1) y (N-k) grados de libertad para un nivel dado (done N= N +…+ Nk y k = número de grupos). Si el valor estadístico F calculado es mayor que (Fcrit) se

concluye que las medias entre los grupos no son iguales. El examen de las medias permite entonces valorar la importancia relativa de cada grupo sobre la medida dependiente. Aunque el contraste F difiere la hipótesis nula de igualdad de las medias, no resuelve la cuestión de qué medidas son diferentes.[24]

1.4 PROGRAMACIÓN NO LINEAL

La programación no lineal es la alternativa que existe cuando los problemas de programación lineal clásica están fuera de alcance, o no se avista una solución mediante los métodos comunes; una forma de identificar un problema de programación no lineal es cuando los modelos de programación no cumplen la condición de que su función objetivo o sus funciones de restricción son lineales. De una manera formal se puede afirmar que el problema de programación no lineal se basa en:

$$
\begin{aligned}
& x = (x_1, x_2, ..., x_n) \\
& \text{para maximizar } f(x), \text{ sujeta a :} \\
& \mathrm{g_i(x)} \le b_i, \qquad \text{para } i = 1, 2, ..., m, \\
& y \\
& x \ge 0
\end{aligned}
\qquad \textbf{(8)}
$$

Donde $f(x)$ y las $\mathrm{g_i(x)}$ son funciones dadas de n variables de decisión. (A efectos prácticos se considera que las funciones son diferenciables o continuas en todas partes, o son funciones lineales por partes). Aunque no existe un algoritmo que resuelva todos los problemas de programación no lineal, muchos problemas específicos se ajustan a este formato.

En algunas circunstancias la no linealidad también puede surgir en las funciones de restricción de $\mathrm{g_i(x)}$ de manera similar como en las funciones objetivo. Se han desarrollado diferentes clases para resolver problemas de programación no lineal, las clases más comunes son: optimización no restringida, optimización linealmente restringida, programación cuadrática, programación convexa, programación no convexa, programación geométrica, programación fraccional, programación separable etc.[25]

---

[24] HAIR. Op Cit. p. 347 - 349

[25] HILLER y LIBERMAN Investigación de operaciones, séptima edición Junio de 2003, editorial McGraw Hill, p. 654, 664 - 669.

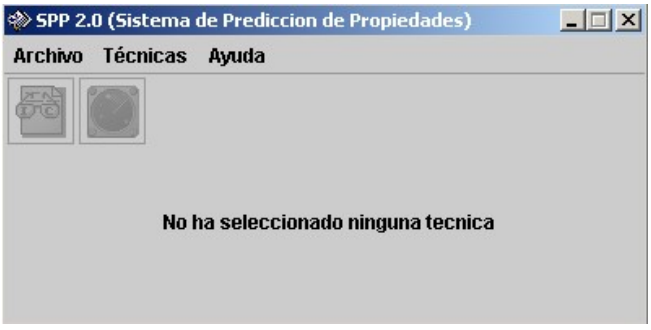
2. DISEÑO DE LA APLICACIÓN

2.1 DESCRIPCIÓN GENERAL

Se hará una breve descripción del prototipo Sistema de Predicción de Propiedades (SPP), diseñado por Ricardo Farfán y Rafael Contreras, así como las modificaciones respectivas hechas ellas por los autores de este documento. El objetivo del prototipo es dar pautas al usuario para analizar, predecir y concluir en qué proporción y cuáles variables independientes ofrecen mejores resultados a la variable dependiente (rendimiento DMO). El prototipo funciona con la carpeta SPP que debe estar localizada en C:/archivos de programa. Esta carpeta contiene la carpeta config (archivos por defecto para consultas a la base de datos), gams (archivo configuración gams), icons (iconos del prototipo), solver (modelo, datos y modelo de respuesta .dat), temp (*dataset*.arff, archivo respuestas globales)

Este prototipo permite hacer consultas a la base de datos Oracle y *Silab*, del Instituto Colombiano de Petróleos-ICP, cargar sábanas de datos (archivos de datos .csv) con el fin de analizar los datos con las técnicas Análisis de Componentes Principales, Análisis de Regresión, Análisis de Cluster, ANOVA, Redes Bayesianas, Árboles de Decisión por medio de *WEKA* (paquete de clases en java) y Programación no lineal a través del software gams (General Algebraic Modeling System). Debido a estos cambios el nombre del prototipo ahora es Sistema de Predicción de Propiedades (SPP 2.0). La ventana principal se muestra en la Figura 10.

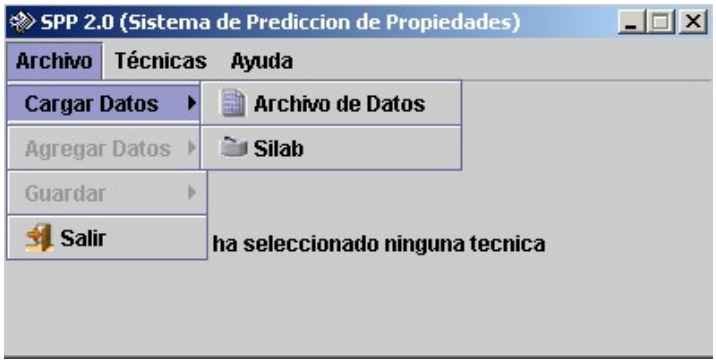
Figura 10. Venta principal SPP 2.0



Fuente: autores del proyecto

El menú archivo en la ventana principal de la Figura 10. Permite cargar datos y agregar datos desde un archivo .csv o desde el *Silab*, guardar los datos y resultados de ejecutar o analizar una técnica del menú Técnicas, como puede verse en la Figura 11.

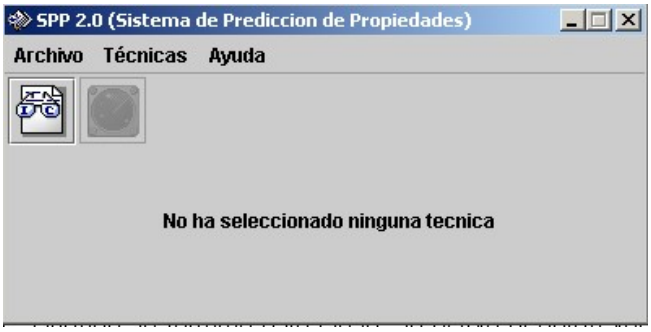
Figura 11. Menú Archivo/Cargar Datos



Fuente: autores del proyecto

La opción Archivo de Datos del menú Archivo/Cargar Datos (Figura 11) permite al usuario buscar una base de datos o *datawarehouse* (bodega donde están almacenados todos los datos necesarios para realizar las funciones de gestión de la empresa, de manera que puedan utilizarse fácilmente según se necesiten) con extensión .csv. Estos datos se cargan en un *dataset* que puede verse oprimiendo el botón “ver datos” como se ve en la Figura 12.

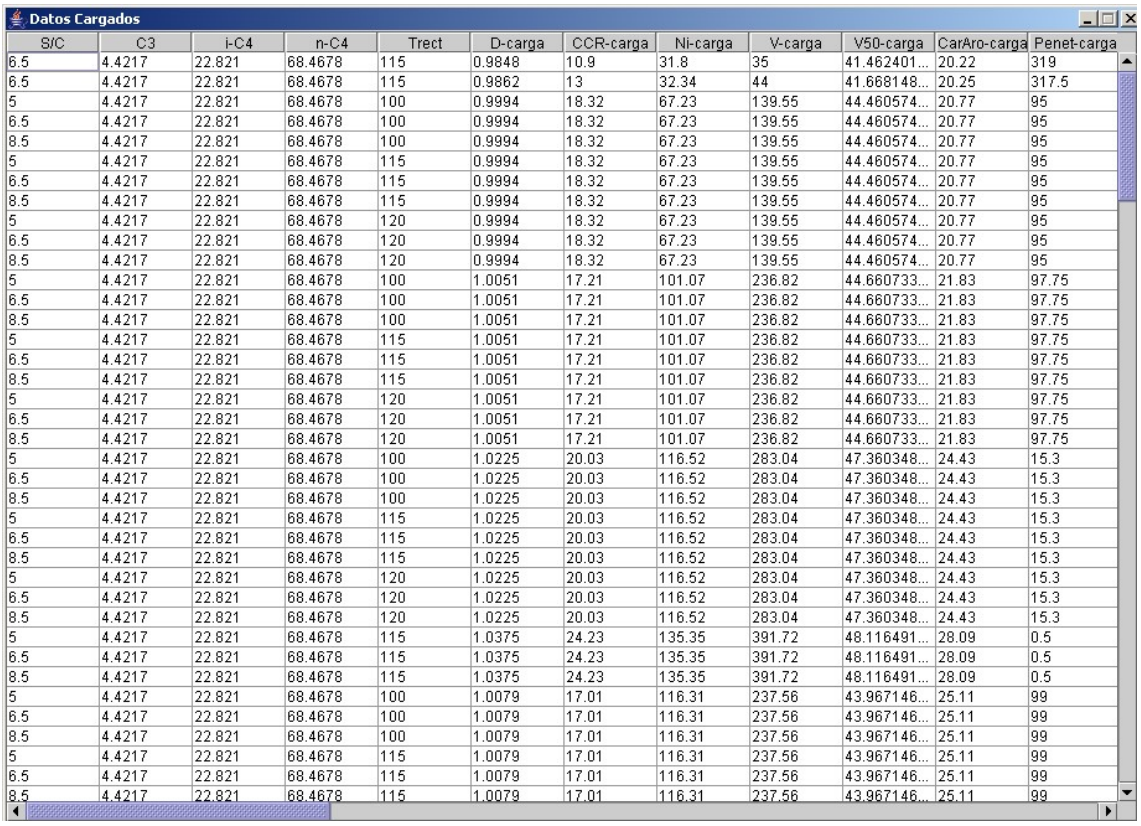
Figura 12. Ver datos



Fuente: autores del proyecto

Si el usuario activa este botón, el prototipo muestra el *dataset* en una Tabla como se puede apreciar en la Figura 13:

Figura 13. Datos Cargados



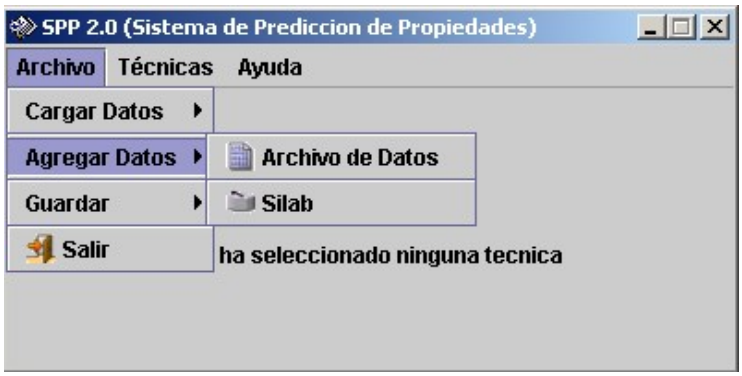
S/C	C3	i-C4	n-C4	Trect	D-carga	CCR-carga	Ni-carga	V-carga	V50-carga	CarAro-carga	Penet-carga
6.5	4.4217	22.821	68.4678	115	0.9848	10.9	31.8	35	41.462401...	20.22	319
6.5	4.4217	22.821	68.4678	115	0.9862	13	32.34	44	41.668148...	20.25	317.5
5	4.4217	22.821	68.4678	100	0.9994	18.32	67.23	139.55	44.460574...	20.77	95
6.5	4.4217	22.821	68.4678	100	0.9994	18.32	67.23	139.55	44.460574...	20.77	95
8.5	4.4217	22.821	68.4678	100	0.9994	18.32	67.23	139.55	44.460574...	20.77	95
5	4.4217	22.821	68.4678	115	0.9994	18.32	67.23	139.55	44.460574...	20.77	95
6.5	4.4217	22.821	68.4678	115	0.9994	18.32	67.23	139.55	44.460574...	20.77	95
8.5	4.4217	22.821	68.4678	115	0.9994	18.32	67.23	139.55	44.460574...	20.77	95
5	4.4217	22.821	68.4678	120	0.9994	18.32	67.23	139.55	44.460574...	20.77	95
6.5	4.4217	22.821	68.4678	120	0.9994	18.32	67.23	139.55	44.460574...	20.77	95
8.5	4.4217	22.821	68.4678	120	0.9994	18.32	67.23	139.55	44.460574...	20.77	95
5	4.4217	22.821	68.4678	100	1.0051	17.21	101.07	236.82	44.660733...	21.83	97.75
6.5	4.4217	22.821	68.4678	100	1.0051	17.21	101.07	236.82	44.660733...	21.83	97.75
8.5	4.4217	22.821	68.4678	100	1.0051	17.21	101.07	236.82	44.660733...	21.83	97.75
5	4.4217	22.821	68.4678	115	1.0051	17.21	101.07	236.82	44.660733...	21.83	97.75
6.5	4.4217	22.821	68.4678	115	1.0051	17.21	101.07	236.82	44.660733...	21.83	97.75
8.5	4.4217	22.821	68.4678	115	1.0051	17.21	101.07	236.82	44.660733...	21.83	97.75
5	4.4217	22.821	68.4678	120	1.0051	17.21	101.07	236.82	44.660733...	21.83	97.75
6.5	4.4217	22.821	68.4678	120	1.0051	17.21	101.07	236.82	44.660733...	21.83	97.75
8.5	4.4217	22.821	68.4678	120	1.0051	17.21	101.07	236.82	44.660733...	21.83	97.75
5	4.4217	22.821	68.4678	100	1.0225	20.03	116.52	283.04	47.360348...	24.43	15.3
6.5	4.4217	22.821	68.4678	100	1.0225	20.03	116.52	283.04	47.360348...	24.43	15.3
8.5	4.4217	22.821	68.4678	100	1.0225	20.03	116.52	283.04	47.360348...	24.43	15.3
5	4.4217	22.821	68.4678	115	1.0225	20.03	116.52	283.04	47.360348...	24.43	15.3
6.5	4.4217	22.821	68.4678	115	1.0225	20.03	116.52	283.04	47.360348...	24.43	15.3
8.5	4.4217	22.821	68.4678	115	1.0225	20.03	116.52	283.04	47.360348...	24.43	15.3
5	4.4217	22.821	68.4678	120	1.0225	20.03	116.52	283.04	47.360348...	24.43	15.3
6.5	4.4217	22.821	68.4678	120	1.0225	20.03	116.52	283.04	47.360348...	24.43	15.3
8.5	4.4217	22.821	68.4678	120	1.0225	20.03	116.52	283.04	47.360348...	24.43	15.3
5	4.4217	22.821	68.4678	115	1.0375	24.23	135.35	391.72	48.116491...	28.09	0.5
6.5	4.4217	22.821	68.4678	115	1.0375	24.23	135.35	391.72	48.116491...	28.09	0.5
8.5	4.4217	22.821	68.4678	115	1.0375	24.23	135.35	391.72	48.116491...	28.09	0.5
5	4.4217	22.821	68.4678	100	1.0079	17.01	116.31	237.56	43.967146...	25.11	99
6.5	4.4217	22.821	68.4678	100	1.0079	17.01	116.31	237.56	43.967146...	25.11	99
8.5	4.4217	22.821	68.4678	100	1.0079	17.01	116.31	237.56	43.967146...	25.11	99
5	4.4217	22.821	68.4678	115	1.0079	17.01	116.31	237.56	43.967146...	25.11	99
6.5	4.4217	22.821	68.4678	115	1.0079	17.01	116.31	237.56	43.967146...	25.11	99
8.5	4.4217	22.821	68.4678	115	1.0079	17.01	116.31	237.56	43.967146...	25.11	99

Fuente: autores del proyecto

La opción *Silab* del menú Archivo/Cargar Datos en la Figura 11 fue modificada, es decir, el SPP permitía consultar elementos de carga y DMO, en SPP 2.0 se consulta además de carga y DMO, demex y solvente. Una vez el usuario escoja esta opción, debe indicar el archivo .csv donde se encuentran los sample id de los elementos a extraer de la base de datos *Silab*. La clase conexión encargada de este evento solicita un archivo donde se encuentran los simple id que corresponden a los alias de cada elemento (Fondo de vacío, DMO, Demex y Solvente) para iniciar la conexión con el *Silab*, internamente se lee el archivo config.txt donde está la ip, el puerto, el nombre de la base de datos, nombre de usuario y la contraseña de usuario. Cuando se termina está clase, se activa el botón *Ver datos* como se ve en la Figura 12 anterior.

Una vez realizada correctamente las dos tareas mencionadas anteriormente, se activan las otras opciones que son A*gregar datos* y G*uardar*, que pueden ser observadas en la Figura 14.

Figura 14. Menú Archivo/Agregar Datos

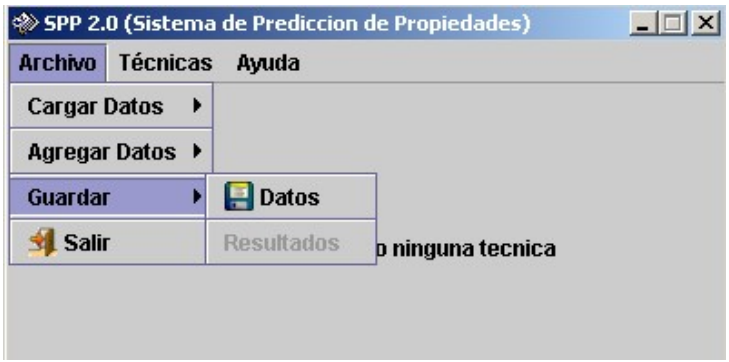


Fuente: autores del proyecto

El submenú *Agregar Datos* de la Figura 14 realiza las mismas funciones que el menú *Cargar Datos* solo que agrega los nuevos datos al final del *dataset* ya creado con anterioridad por el usuario.

La opción *Datos* de menú Archivo/Guardar solicita al usuario la ruta y nombre del archivo .csv para guardar el *dataset* con los datos cargados y/o agregados con anterioridad. Esta opción se ve en la Figura 15.

Figura 15. Menú Archivo/Guardar



Fuente: autores del proyecto

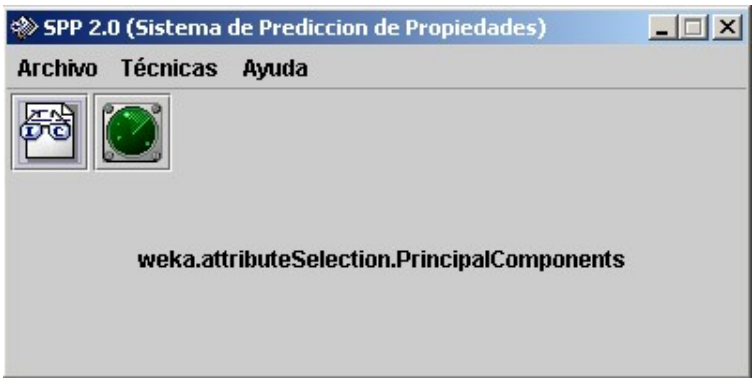
El menú *Técnicas* de la Figura 16 es el encargado de llevar a cabo la(s) técnica(s) seleccionada(s) por el usuario así como mostrar los resultados respectivos.

Figura 16. Menú Técnicas

Fuente: autores del proyecto

Cuando el usuario selecciona una técnica el botón *Analizar*, se activa (ver Figura 17).

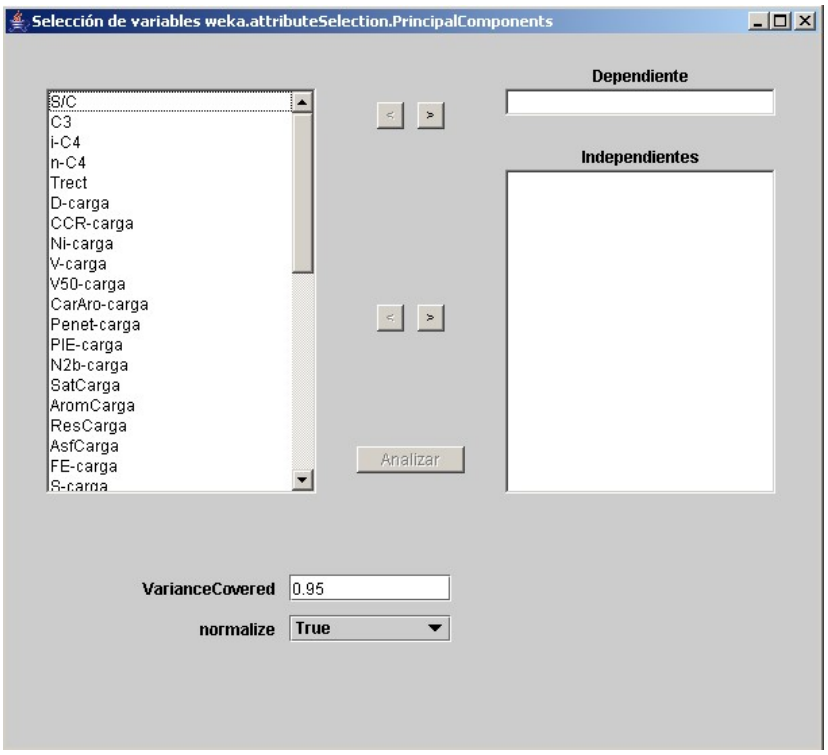
Figura 17. Botón Analizar



Fuente: autores del proyecto

Y este a su vez ejecuta una nueva ventana donde el usuario seleccionará la variable dependiente e independiente(s), así como las opciones propias de cada una de las técnicas. (Ver Figura 18).

Figura 18. Seleccionar variables



Fuente: autores del proyecto

Si el usuario selecciona la técnica ANOVA (análisis de varianza), se carga un *dataset* con las variables a analizar y el prototipo establece comunicación con la clase ANOVA, donde se calculan las medias por filas y columnas y la media de las medias, la suma total de cuadrados (STT), la suma de cuadrados del tratamiento – intermuestral (SSTR), la suma de cuadrados del error o intramuestral (SEE) y determina cuáles variables son o no significativas, y se muestran en la interfaz de respuesta.

Para el caso de Gams, se escriben las variables a analizar en un archivo .gms y el directorio por *default* (SPP/solver), en la clase gams y se da paso a la clase llamadogams el cual estable la conexión con el gams.exe enviándole como parámetros los archivos .gms de los datos y el modelo. Cuando este sistema termina los cálculos retorna un modelo .dat y la clase llamadogams lo lee y se imprime en la interfaz de repuesta.

Para la técnicas *cluster* (algoritmo *simplekmedias*), se cargan las variables en un *dataset* en formato *arff* de *Weka* (ver Tabla 2), se hace comunicación con la clase

cluster y esta a su vez con el paquete de clases *Weka* y se retorna el resultado en la interfaz de respuesta.

Tabla 2. Archivo A*rff Weka*

@relation nombre _ archivo
@attribute nombre _ variable {valores que contiene}
@data valores de todas las variables seguidas por comas

```
@relation weather
@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
Zuñí,85,85,FALSE,no
Zuñí,80,90,Trueno
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
Zuñí,72,95,FALSE,no
Zuñí,69,70,FALSE,yes
rainy,75,80,FALSE,yes
Zuñí,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Fuente: Software de Minería de Datos *Weka*

Para la técnica Redes de Bayes (algoritmo *bayes net*), se cargan las variables en un *dataset* en formato *Arff* de *Weka*, se hace comunicación con la clase clasificadores y esta a su vez con el paquete de clases W*eka* y se retorna el resultado en la interfaz de respuesta.

A continuación se mostrará cómo y por qué fueron modificadas algunas técnicas:

En los ejemplos que se presentan en los siguientes ítems se usaron las variables independientes (ver Tabla 3) y la variable dependiente, RendDMO; (ver Tabla 4).

Tabla 3. Variables independientes

NOMBRE VARIABLE	ESCALA	NOMBRE VARIABLE	ESCALA
CCR-carga	Numérica	ResCarga	Numérica
Ni-carga	Numérica	AsfCarga	Numérica
D-carga	Numérica	FE-carga	Numérica
V-carga	Numérica	S-carga	Numérica
V50-carga	Numérica	Trect	Numérica
S/C	Numérica	CarAro-carga	Numérica
Penet-carga	Numérica	AromCarga	Numérica
PIE-carga	Numérica	T10	Numérica
N2b-carga	Numérica	T30	Numérica
SatCarga	Numérica	T50	Numérica
BCMI-carga	Numérica	C3	Numérica
i-c4	Numérica	n-c4	Numérica

Fuente: ICP

Tabla 4. Variable dependiente

NOMBRE VARIABLE	ESCALA
RendDMO	Numérica
CatRendDMO	Categórica o nominal

Fuente: ICP

2.1.1 Nueva representación en la técnica análisis de regresión es controlada por la clase Clasificadores, una vez se envíen a W*eka* el nombre de la técnica, las opciones y el *dataset,* se obtiene el modelo lineal presentado en columnas como se ve en la Figura 19. Para el análisis de esta técnica la variable dependiente debe ser numérica.

Figura 19. Modelo lineal por *Weka*

```
Linear Regression Model
----------------------
RendvDMO =
      1.6791 * S/C +
   -125.9436 * D-carga +
     -0.6604 * CCR-carga +
     -0.1343 * Ni-carga +
     -0.0192 * V-carga +
     -0.0386 * PIE-carga +
```

```
-0.3741 * ResCarga +
245.3181
```

Fuente: Sistema de Predicción de Propiedades, SPP 2.0

Debido a este formato, se cambia el modelo de la Ecuación lineal como se muestra en la Figura 20.

Figura 20. Modificación del modelo lineal

```
Linear Regression Model
----------------------
RendvDMO = 1.6791 * S/C - 125.9436 * D-carga - 0.6604 * CCR-carga - 0.1343 * Ni-carga
- 0.0192 * V-carga - 0.0386 * PIE-carga - 0.3741 * ResCarga + 245.3181
```

Fuente: Sistema de Predicción de Propiedades, SPP 2.0

Para este cambio se crearon las siguientes líneas de código:

```
auxCoeflinea = auxCoeflinea.replaceAll("\n","!");
auxCoeflinea = auxCoeflinea.replaceAll("  ","");
String auxCoeflinea1="";
char VauxCoef1[] = auxCoeflinea.toCharArray();
for(int i=0; i<VauxCoef1.length;i++){
int j=i+1;
if(VauxCoef1[i]=='-' && (VauxCoef1[j]=='0' || VauxCoef1[j]=='1' || VauxCoef1[j]=='2' || VauxCoef1[j]=='3' ||
VauxCoef1[j]=='4' || VauxCoef1[j]=='5' || VauxCoef1[j]=='6' || VauxCoef1[j]=='7' || VauxCoef1[j]=='8' ||
VauxCoef1[j]=='9')){
VauxCoef1[i-2]='-';
VauxCoef1[i]=' ';  }
}
for(int i=0; i<VauxCoef1.length;i++){
auxCoeflinea1 = auxCoeflinea1 + VauxCoef1[i];
if(i==175){
auxCoeflinea1 = auxCoeflinea1 + "\n\t"; }
}
auxCoeflinea1= auxCoeflinea1.replaceAll("!"," ");
auxCoeflinea1= auxCoeflinea1.replaceAll("  "," ");
```

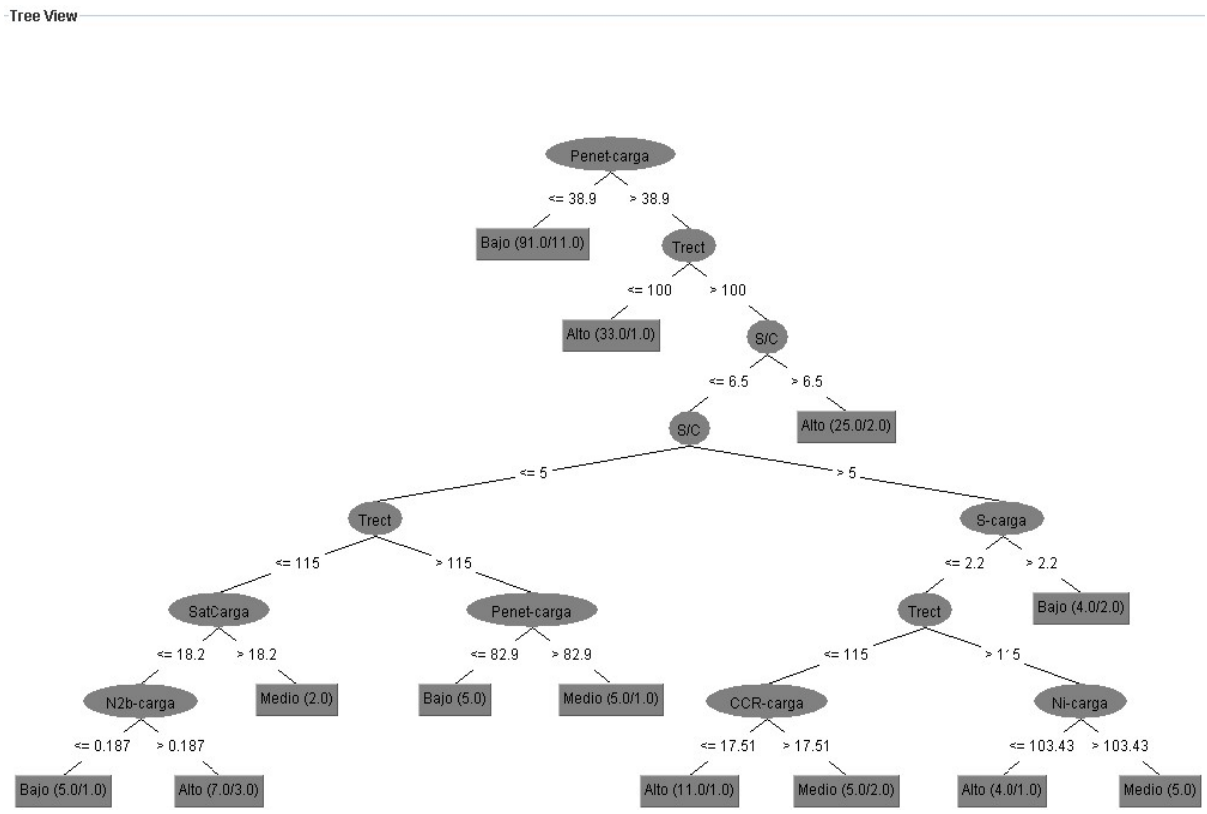
2.1.2 Árbol gráfico en la técnica árboles de decisión la técnica Árboles de Decisión (algoritmo j48) es controlada por la clase Clasificadores, una vez se envíe a *Weka* el nombre de la técnica, las opciones y el archivo *Arff*, se obtienen los resultados que se presentan en la interfaz de respuesta. Se incluyen unas líneas de código para generar el Árbol Gráfico ya que es más fácil de interpretar que el plano que se muestra en la Figura 21. Para el análisis de esta técnica se usa la variable dependiente categórica, cartRendDMO.

Figura 21. Resultados técnica Árboles de Decisión

```
J48 unpruned tree
------------------

Penet-carga <= 38.9
|   Trect <= 100
|   |   V-carga <= 235.38
|   |   |   S/C <= 5: Bajo (4.0/2.0)
|   |   |   S/C > 5: Alto (8.0/4.0)
|   |   V-carga > 235.38: Bajo (9.0)
|   Trect > 100: Bajo (70.0/4.0)
Penet-carga > 38.9
|   Trect <= 100: Alto (33.0/1.0)
|   Trect > 100
|   |   S/C <= 6.5
|   |   |   S/C <= 5
|   |   |   |   Trect <= 115
|   |   |   |   |   SatCarga <= 18.2
|   |   |   |   |   |   N2b-carga <= 0.187: Bajo (5.0/1.0)
|   |   |   |   |   |   N2b-carga > 0.187: Alto (7.0/3.0)
|   |   |   |   |   SatCarga > 18.2: Medio (2.0)
|   |   |   |   Trect > 115
|   |   |   |   |   Penet-carga <= 82.9: Bajo (5.0)
|   |   |   |   |   Penet-carga > 82.9: Medio (5.0/1.0)
|   |   |   S/C > 5
|   |   |   |   S-carga <= 2.2
|   |   |   |   |   Trect <= 115
|   |   |   |   |   |   CCR-carga <= 17.51: Alto (11.0/1.0)
|   |   |   |   |   |   CCR-carga > 17.51: Medio (5.0/2.0)
|   |   |   |   |   Trect > 115
|   |   |   |   |   |   Ni-carga <= 103.43: Alto (4.0/1.0)
|   |   |   |   |   |   Ni-carga > 103.43: Medio (5.0)
|   |   |   |   S-carga > 2.2: Bajo (4.0/2.0)
|   |   S/C > 6.5: Alto (25.0/2.0)
Number of Leaves  :    16
Size of the tree : 31

=== Summary ===
Correctly Classified Instances         178               88.1188 %
Incorrectly Classified Instances        24               11.8812 %
Kappa statistic                          0.7987
Mean absolute error                      0.1166
Root mean squared error                  0.2415
Root relative squared error             53.8417 %
Total Number of Instances              202

=== Confusion Matrix ===
  a  b  c   <-- classified as
 76  1  2 |  a = Alto
  7 88  1 |  b = Bajo
  5  8 14 |  c = Medio
```

Fuente: Sistema de Predicción de Propiedades, SPP 2.0

Figura 22. Árbol gráfico



Fuente: Sistema de Predicción de Propiedades, SPP 2.0

Las líneas de código para generar el Árbol Gráfico son:

```
J48 cls = new J48();
instances.setClassIndex(instances.numAttributes() - 1);
cls.buildClassifier(instances);
javax.swing.JFrame jf = new javax.swing.JFrame("Weka Classifier Tree Visualizer: J48");
jf.setSize(500,400);
jf.setBounds(100, 100 , 500, 400);
java.awt.BorderLayout ley = new java.awt.BorderLayout();
jf.getContentPane().setLayout(ley);
TreeVisualizer tv = new TreeVisualizer(null,cls.graph(),new PlaceNode2());
jf.getContentPane().add(tv, ley.CENTER);
jf.setVisible(true);
tv.fitToScreen();
```

Fuente: autores del proyecto

2.1.3 Nuevos métodos en la técnica componentes principales es controlada por la clase Principal, una vez se envíe a *Weka* el nombre de la

técnica, las opciones y el archivo *Arff* a la, se obtiene la matriz de correlación que permite realizar nuevos métodos para calcular propiedades que permiten hacer más clara la interpretación de los resultados de la técnica para determinar cuales variables afectan la correlación y eliminarla del análisis.

En las reuniones que se realizaron con el usuario se presentaron falencias o ausencias de algunas funciones que facilitan la interpretación de resultados. Entre estos se encuentran la Matriz de Correlación Anti-imagen, el Índice KMO (Kaiser-Meyer-Olkin) de Adecuación de la Muestra y el Test de esfericidad de Bartlett en la técnica de componentes principales.

2.1.3.1 Determinante de la matriz de correlación un determinante muy bajo indicará altas inter-correlaciones entre las variables, pero no debe ser cero (matriz no singular), pues esto indicaría que algunas de las variables son linealmente dependientes y no se podrían realizar ciertos cálculos necesarios en el Análisis Factoria). El resultado del determinante está en la Figura 23.

Figura 23. Determinante

```
Correlation matrix
 1     0.87  0.68  0.63  0.55  0     -0.72  0.2   0.37 -0.72 -0.2   0.75  0.16  0.49
 0.87  1     0.59  0.54  0.57  0.01 -0.79  0.32  0.35 -0.73 -0.09  0.82  0.05  0.33
 0.68  0.59  1     0.6   0.43 -0.01 -0.67 -0.15  0.37 -0.52 -0.41  0.65  0.11  0.53
 0.63  0.54  0.6   1     0.44  0    -0.56  0.14  0.29 -0.33 -0.19  0.64 -0.06  0.53
 0.55  0.57  0.43  0.44  1     0.01 -0.59  0.21  0.26 -0.52  0.01  0.43  0.17  0.43
 0     0.01 -0.01  0     0.01  1    -0.01  0.02 -0.02  0     0     0.01 -0.02  0.02
-0.72 -0.79 -0.67 -0.56 -0.59 -0.01  1    -0.24 -0.54  0.75 -0.07 -0.6  -0.19 -0.54
 0.2   0.32 -0.15  0.14  0.21  0.02 -0.24  1     0.25 -0.49  0.54  0     0.44  0.23
 0.37  0.35  0.37  0.29  0.26 -0.02 -0.54  0.25  1    -0.57  0.23  0.2   0.42  0.28
-0.72 -0.73 -0.52 -0.33 -0.52  0     0.75 -0.49 -0.57  1    -0.28 -0.38 -0.69 -0.46
-0.2  -0.09 -0.41 -0.19  0.01  0    -0.07  0.54  0.23 -0.28  1    -0.46  0.5   0.03
 0.75  0.82  0.65  0.64  0.43  0.01 -0.6   0     0.2  -0.38 -0.46  1    -0.29  0.31
 0.16  0.05  0.11 -0.06  0.17 -0.02 -0.19  0.44  0.42 -0.69  0.5  -0.29  1     0.28
 0.49  0.33  0.53  0.53  0.43  0.02 -0.54  0.23  0.28 -0.46  0.03  0.31  0.28  1

EL DETERMINANTE DE LA MATRIZ ES:
---------------------------------------------------------
7.558424958205402E-7
```

Fuente: Sistema de Predicción de Propiedades, SPP 2.0

2.1.3.2 Test de esfericidad de Bartlett este valor se puede observar implementando el análisis factorial en el prototipo, y los resultados se aprecian en la Figura 24. Se debe suponer la normalidad entre variables y hacer contraste de las correlaciones de las variables; si el contraste no es adecuado significa que tal vez el tamaño de la muestra no es suficiente (se necesita una muestra más

grande), entonces no se recomienda aplicar análisis factorial ya que las variables no están bien correlacionadas.[26]

Figura 24. Resultado del Test de esfericidad

COEFICIENTE DE ESFERICIDAD DE BARTLETT
2804.9911314454966

Fuente: autores, Spp 2.0

El Test de esfericidad de Bartlett se utiliza para probar la hipótesis nula que afirma que las variables no están correlacionadas en la población; es decir, comprueba si la matriz de correlaciones es una matriz de identidad. Se puede dar como válidos aquellos resultados que presenten un valor elevado del test y cuya fiabilidad sea baja, en este caso se rechaza la hipótesis nula y se continúa con el Análisis[27].

La Ecuación del Test de esfericidad se muestra en la Ecuación 9:

$$\mathbf{X^2} = - (n - 1\ (2p + 5)/6)\ \ln |R^*| \ \mathbf{(9)}$$

Bajo la hipótesis nula resulta $\mathbf{X^2_{(p2-p)/2}}$

Donde:

p es el número de variables

|R*| es el determinante de la matriz de correlaciones normales

2.1.3.3 Índice KMO (Kaiser-Meyer-Olkin) de adecuación de la muestra La Ecuación 10 muestra el cálculo del KMO.

[26] Estadístico, es el sitio Web especialistas en consultoría y formación estadística, integrado por expertos en los programas SPSS, SAS, CLEMENTINE entre otros, en la Web de data Mining Institute encontrará todo lo referente a la estadística: cursos, artículos, software, enlaces, consultoría, libros, diccionario estadístico y tests. [Online, Artículo estadístico] 2004. [Citado el 07 de Febrero 2006] Disponible en <http://www.estadistico.com/arts.htm?20001106>

[27] Estadístico, es el sitio Web especialistas en consultoría y formación estadística, integrado por expertos en los programas SPSS, SAS, CLEMENTINE entre otros, en la Web de data Mining Institute encontrará todo lo referente a la estadística: cursos, artículos, software, enlaces, consultoría, libros, diccionario estadístico y tests. [Online, Artículo estadístico] 2004. [Citado el 25 de agosto de 2006] <http://www.estadistico.com/dic.html?p=4135&PHPSESSID= 83d26dfa82 897dc24a9ec5c8225dd61a>

$$KMO = \frac{\sum\sum_{i \neq j} r^2{}_{ji}}{\sum\sum_{i \neq j} r^2{}_{ji} + \sum\sum_{i \neq j} a^2{}_{ji}} \quad \textbf{(10)}$$

Donde:

rji es el coeficiente de correlación observada en la matriz de correlación normal.

aji es el coeficiente de correlación parcial en la matriz de correlaciones parciales.

Los coeficientes están midiendo la correlación existente entre las variables, teniendo en cuenta que se elimina la influencia del resto sobre ellas. Estas influencias se pueden interpretar como los efectos correspondientes a los factores comunes; y por tanto, al eliminarlos, *aji* representará la correlación entre los factores únicos de las dos variables, que teóricamente tendría que ser nula. Si hubiese correlación entre las variables, estos coeficientes deberían estar próximos a 0, lo que arrojaría un KMO próximo a 1. Por el contrario, valores del KMO próximos a 0 retraerían la idea de aplicar un análisis factorial[28]

Está comúnmente aceptado que:

- Si KMO < 0.5 no resultaría aceptable para hacer un ACP
- Si 0.5 < KMO < 0.6 grado de correlación medio, y habría aceptación media
- Si KMO > 0.7 indica alta correlación y, por tanto, conveniencia de ACP

El resultado del *KMO* en el ejemplo con las variables independientes y la variable dependiente numérica RendvDMO indica que es conveniente continuar con el análisis. Esto puede verse en la Figura 25.

Figura 25. Resultado índice de KMO

[28] Estadístico, es el sitio Web especialistas en consultoría y formación estadística, integrado por expertos en los programas SPSS, SAS, CLEMENTINE entre otros, en la Web de data Mining Institute encontrará todo lo referente a la estadística: cursos, artículos, software, enlaces, consultoría, libros, diccionario estadístico y tests. [Online, Artículo estadístico] 2004. [Citado el 07 de Febrero 2006] Disponible en <http://www.estadistico.com/arts.htm?20001106>

0.745866743041809

Fuente: autores, spp 2.0

2.1.3.4 Matriz de Correlación anti-imagen en el análisis de compones principales (ACP) se ejecuta una análisis de correlaciones analizando la muestra la cual cuenta con **n** registros y un número determinado de variables; partiendo de esta matriz de correlaciones se puede hacer una aproximación de carácter intuitivo para poder saber cuales son las variables que están afectado claramente al modelo que se está ejecutando, ya que su correlación con el sistema propio es muy bajo (no están fuertemente correlacionadas al sistema), de esta manera se pueden descartar ciertas variables. Para poder ejecutar este análisis es necesario preparar la matriz anti-imagen de correlaciones parciales (ver Figura 22), que permite detectar cuales son las correlaciones bajas en el sistema, analizando su diagonal, la cual indica que si los coeficientes de la variable son menores de 0,3 (para este caso específico, resultado de un análisis de correlaciones enfocado a procesos de Minería de Datos) deben ser omitidas estas variables del sistema, de lo contrario la correlación es fuerte para poder contar con estas.

Los cálculos para realizar la matriz anti-imagen de correlaciones parciales se mostrarán en el siguiente ejemplo (está definido en Figuras hechas totalmente del autor en base a la Ecuación, matriz de datos y SPSS citado más adelante), donde se resolverá un sistema de 3x3 para hacerlo sencillo, y después se mostrará una regla general para dimensiones más grandes de matrices de correlación normal. Se tiene el siguiente sistema de 3X3 con un n=5, y se harán los cálculos de la matriz anti-imagen de correlaciones parciales, partiendo de la matriz de correlaciones calculada con anterioridad.

Tabla 5. Muestra estadística

Datos (N=5)

Rendim	Ansied	Neurot
9	3	5
3	12	15
6	8	8
2	9	7
7	7	6

Fuente: PEREA, Manuel. Associate Professor Universidad de Valencia. Bloque III. Caracterización de la relación entre variables. 2006 [online, Presentacion power point]. [Citado el 30 de agosto 2006] Disponible en Internet: <http://www.uv.es/~mperea/T8_APD.ppt>

Con la muestra anterior se está en la capacidad de calcular la matriz de correlaciones, para esto se hace uso del SPSS 12.0.

Tabla 6. Matriz de correlaciones

Matriz de correlaciones

		Rendim	Ansied	Neurot
Correlación	Rendim	1.000	-.865	-.600
	Ansied	-.865	1.000	.853
	Neurot	-.600	.853	1.000

Fuente: SPSS 12.0.

Partiendo de esta matriz, se realizan los cálculos para hallar los coeficientes secundarios de la matriz anti-imagen de correlaciones parciales; el adjetivo (parciales) viene de que estas correlaciones se calculan entre dos variables (nunca las mismas), excluyendo a una tercera, eso da el carácter de parcialidad en el contexto.

(Todas las gráficas que siguen a continuación para explicar el procedimiento son hechas por los autores a menos que se indique lo contrario).

Los coeficientes contienen la siguiente notación:

r12.3: Correlación parcial de las variables 1 y 2 excluyendo a la variable 3.
r13.2: Correlación parcial de las variables 1 y 3 excluyendo a la variable 2.
r23.1: Correlación parcial de las variables 2 y 3 excluyendo a la variable 1.

Los cuales quedan organizados en la matriz de correlación anti imagen, secuenciados por los dos primeros subíndices, como se muestra continuación.

		Rendim	Ansied	Neurot
Correlación anti-imagen	Rendim		$r_{12.3}$	$r_{13.2}$
	Ansied			$r_{23.1}$
	Neurot			
a Medida de adecuación muestral				

Como la matriz de correlaciones normales y la matriz anti-imagen de correlaciones parciales, son matrices simétricas entonces se tiene:

		Rendim	Ansied	Neurot
Correlación anti-imagen	Rendim		$r_{12.3}$	$r_{13.2}$
	Ansied	$r_{12.3}$		$r_{23.1}$
	Neurot	$r_{13.2}$	$r_{23.1}$	
a Medida de adecuación muestral				

Ahora se muestra cómo realizar los cálculos correspondientes a todos los **r** que se encuentra en la matriz anti-imagen de correlaciones en la Ecuación 11.

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} \quad \textbf{(11)}$$

Fuente: Ibíd., p. 53

Reemplazando en la Ecuación 11 por los valores de la matriz de correlaciones, se tiene:

Matriz de correlaciones

		Rendim	Ansied	Neurot
Correlación	Rendim	1.000	-.865	-.600
	Ansied	-.865	1.000	.853
	Neurot	-.600	.853	1.000

r_{12} r_{13} r_{23}

$$r_{12.3} = \frac{(-0{,}865) - (-0.600) \times (0.853)}{\sqrt{1 - (-0.600)^2}\sqrt{1 - (0.853)^2}}$$

El resultado es:

$$r_{12.3} = -0.845$$

Se multiplica por (-1) y se ubica de la siguiente manera en la matriz.

		Rendim	Ansied	Neurot
Correlación anti-imagen	Rendim		0.845	$r_{13.2}$
	Ansied	0.845		$r_{23.1}$
	Neurot	$r_{13.2}$	$r_{23.1}$	

a Medida de adecuación muestral

Ahora se repite el proceso con la Ecuación 12 para *r13.2*:

$$r_{13.2} = \frac{r_{13} - r_{12} \cdot r_{32}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{32}^2}} \qquad \textbf{(12)}$$

Reemplazando en la Ecuación 12 con los valores de la matriz de correlaciones se tiene:

Matriz de correlaciones

		Rendim	Ansied	Neurot
Correlación	Rendim	1.000	-.865	-.600
	Ansied	-.865	1.000	.853
	Neurot	-.600	.853	1.000

r_{13} r_{12} r_{32}

$$r_{13.2} = \frac{-(0.600) - (-0.865) \times (0.853)}{\sqrt{1 - (-0.865)^2}\sqrt{1 - (0.853)^2}}$$

El resultado es:

$$r_{13.2} = 0.523$$

Se multiplica por (-1) y se ubica de la siguiente manera en la matriz.

		Rendim	Ansied	Neurot
Correlación anti-imagen	Rendim		0.845	-0.523
	Ansied	0.845		$r_{23.1}$
	Neurot	-0.523	$r_{23.1}$	
a Medida de adecuación muestral				

Para terminar con la Ecuación 13 se calcula el último coeficiente secundario, *r23.1*

$$r_{23.1} = \frac{r_{23} - r_{21} \cdot r_{31}}{\sqrt{1 - r_{21}^2}\sqrt{1 - r_{31}^2}} \qquad \textbf{(13)}$$

Reemplazando en la Ecuación 13 con los valores de la matriz de correlaciones, se tiene:

Matriz de correlaciones

		Rendim	Ansied	Neurot
Correlación	Rendim	1.000	-.865	-.600
	Ansied	-.865	1.000	.853
	Neurot	-.600	.853	1.000

r_{21} r_{23} r_{31}

$$r_{23.1} = \frac{(0.853) - ((-0.865) \times (-0.600))}{\sqrt{1 - (-0.865)^2}\sqrt{1 - (-0.600)^2}}$$

El resultado es:

$$r_{23.1} = 0.831$$

Se multiplica por (-1) y se ubica de la siguiente manera en la matriz.

		Rendim	Ansied	Neurot
Correlación anti-imagen	Rendim		0.845	-0.523
	Ansied	0.845		-0.831
	Neurot	-0.523	-0.831	
a Medida de adecuación muestral				

Finalmente se obtienen todos los coeficientes secundarios de la matriz anti-imagen de correlaciones, estos coeficientes se usarán en el calculo de los coeficientes de correlación parcial que se ubican en la diagonal principal de esta matriz, y que son los que determinan si las variables están bien correlacionadas al sistema (si el coeficiente es mayor de 0,3 en el caso particular). Para calcular los coeficientes de correlación parcial que se ubican en la matriz anti-imagen de correlaciones parciales, se usa la Ecuación 14.

$$MSA_i = \frac{\sum_{j \neq i} r_{ij}^2}{\sum_{j \neq i} r_{ij}^2 + \sum_{j \neq i} a_{ij}^2} \quad \textbf{(14)}$$

Fuente: Estadístico, es el sitio Web especialistas en consultoría y formación estadística, integrado por expertos en los programas SPSS, SAS, CLEMENTINE entre otros, en la Web de data Mining Institute encontrará todo lo referente a la estadística: cursos, artículos, software, enlaces, consultoría, libros, diccionario estadístico y tests. [Online, Artículo estadístico] 2004. [Citado el 07 de Febrero 2006] Disponible en <http://www.estadistico.com/arts.htm?20001106>

Donde *r* se interpreta como los coeficientes de la matriz de correlaciones, y *a* como los coeficientes parciales de la matriz anti-imagen de correlaciones parciales.

Reemplazando en la Ecuación 14 se calcula el primer elemento para MSA1

i = 1;

$$MSA_1 = \frac{\sum_{j \neq 1} r_{1j}^2}{\sum_{j \neq 1} r_{1j}^2 + \sum_{j \neq 1} a_{1j}^2}$$

$$\sum_{j \neq 1} r_{1j}^2 = (-0.865)^2 + (-0.600)^2 = 1.108225$$

$$\sum_{j \neq 1} a_{1j}^2 = (0.845)^2 + (-0.523)^2 = 0.987554$$

$$MSA_1 = \frac{1.108225}{1.108225 + 0.987554} = 0.528(a)$$

Se ubica el valor de la correlación parcial en la diagonal de la matriz anti-imagen de correlaciones parciales.

		Rendim	Ansied	Neurot
Correlación anti-imagen	Rendim	0.528a	0.845	-0.523
	Ansied	0.845		-0.831
	Neurot	-0.523	-0.831	

a Medida de adecuación muestral

Reemplazando en la Ecuación 14 se calcula el primer elemento para MSA2

i = 2;

$$MSA_2 = \frac{\sum_{j \neq 2} r_{2j}^2}{\sum_{j \neq 2} r_{2j}^2 + \sum_{j \neq 2} a_{2j}^2}$$

$$\sum_{j\neq 2} r_{2j}^2 = (-0.865)^2 + (0.853)^2 = 1.4758$$

$$\sum_{j\neq 2} a_{2j}^2 = (0.845)^2 + (0.831)^2 = 1.404586$$

$$MSA_2 = \frac{1.4758}{1.4758 + 1.404586} = 0.512(a)$$

		Rendim	Ansied	Neurot
Correlación anti-imagen	Rendim	0.528a	0.845	-0.523
	Ansied	0.845	0.512a	-0.831
	Neurot	-0.523	-0.831	
a Medida de adecuación muestral				

Finalmente se calcula el tercer y último elemento para MSA3 reemplazando en la Ecuación 14.

i = 3;

$$MSA_3 = \frac{\sum_{j\neq 3} r_{3j}^2}{\sum_{j\neq 3} r_{3j}^2 + \sum_{j\neq 3} a_{3j}^2}$$

$$\sum_{j\neq 3} r_{3j}^2 = (-0.600)^2 + (0.853)^2 = 1.087609$$

$$\sum_{j\neq 3} a_{3j}^2 = (-0.523)^2 + (-0.831)^2 = 0.96409$$

$$MSA_3 = \frac{1.087609}{1.087609 + 0.96409} = 0.530(a)$$

		Rendim	Ansied	Neurot
Correlación anti-imagen	Rendim	0.528a	0.845	-0.523
	Ansied	0.845	0.512a	-0.831
	Neurot	-0.523	-0.831	0.530a
a Medida de adecuación muestral				

Se tiene totalmente construida la matriz anti-imagen de correlaciones parciales, y se observa que los valores de la diagonal principal son todos mayores de 0.3, lo cual indica que todas las variables del sistema están bien correlacionadas entre sí, y se puede contar con todas ellas en los análisis posteriores, y no significan error en los mismos; de lo contrario, sería necesario sacar las variables las cuales tienen correlación menor de 0.3 y correr el análisis nuevamente, para poder analizar si el sistema mejora, o hay otras variables que lo hacen inestable.

En el ejercicio anterior se pudo observar la construcción completa de la matriz anti-imagen de correlaciones parciales, para un sistema de matrices 3X3 con 3 variables, pero si el sistema contiene más de 3 variables el análisis de la matriz aumenta de complejidad considerablemente, ya que los siguientes grados de **rn** (matrices de dimensiones mayores a 3) dependen inmediatamente del anterior r(n-1). Por ejemplo, para un sistema r7 (matriz de dimensiones 7X7), se hará necesario la construcción de r6, r5, r4, r3; cuando ocurre esto el cálculo de las correlaciones parciales secundarias (las que no se encuentran en la diagonal principal de la matriz anti-imagen de correlaciones parciales) se calculan mediante la Ecuación 15.

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{1 - r_{14.3}^2}\sqrt{1 - r_{24.3}^2}} \quad \textbf{(15)}$$

Donde se puede ver claramente que se tiene en cuenta de igual manera la correlación de 2 variables (1 y 2), pero excluyendo las otras 2 del sistema (3 y 4), para un sistema r4, donde se considera de segundo orden, si se observa en la Ecuación 12 el coeficiente que está en rojo, se puede ver la relación que este sistema encierra directamente con el sistema anterior r3, de primer orden.

En la Figura 25 se puede apreciar la Matriz de Correlación Anti-imagen.

Figura 25. Matriz anti-imagen de correlación

MATRIZ ANTI-IMAGEN DE CORRELACIONES

D-carga	CCR-carga	Ni-carga	V-carga	V50-carga	S/C	Penet-carga	PIE-carga	N2b-carga	SatCarga	ResCarga	AsfCarga	FE-carga	S-carga
0.833(a)	-0.374	0.106	-0.405	0.032	0.016	-0.281	0.294	-0.076	0.241	0.308	0.175	0.099	-0.339
-0.374	0.73(a)	-0.069	0.031	-0.182	0.018	-0.0070	-0.338	0.22	0.599	-0.171	-0.45	0.576	0.41
0.106	-0.069	0.832(a)	-0.264	0.084	0.014	0.223	0.413	-0.077	0.051	0.464	-0.069	-0.124	-0.257
-0.405	0.031	-0.264	0.745(a)	-0.105	-0.0010	0.293	-0.309	-0.056	-0.398	-0.23	-0.36	-0.275	-0.057
0.032	-0.182	0.084	-0.105	0.94(a)	-0.0040	0.144	0.08	0.063	-0.061	0.062	0.075	-0.1	-0.166
0.016	0.018	0.014	-0.0010	-0.0040	0.299(a)	-0.0080	-0.014	0.018	0.034	0.012	0.0020	0.037	-0.024
-0.281	-0.0070	0.223	0.293	0.144	-0.0080	0.768(a)	-0.195	0.258	-0.568	0.143	-0.254	-0.58	0.275
0.294	-0.338	0.413	-0.309	0.08	-0.014	-0.195	0.627(a)	-0.082	0.073	-0.06	0.151	-0.061	-0.337
-0.076	0.22	-0.077	-0.056	0.063	0.018	0.258	-0.082	0.867(a)	0.077	-0.126	-0.165	-0.026	0.222
0.241	0.599	0.051	-0.398	-0.061	0.034	-0.568	0.073	0.077	0.675(a)	0.094	0.116	0.925	0.021
0.308	-0.171	0.464	-0.23	0.062	0.012	0.143	-0.06	-0.126	0.094	0.656(a)	0.357	-0.027	-0.148
0.175	-0.45	-0.069	-0.36	0.075	0.0020	-0.254	0.151	-0.165	0.116	0.357	0.814(a)	0.228	-0.212
0.099	0.576	-0.124	-0.275	-0.1	0.037	-0.58	-0.061	-0.026	0.925	-0.027	0.228	0.446(a)	0.0060
-0.339	0.41	-0.257	-0.057	-0.166	-0.024	0.275	-0.337	0.222	0.021	-0.148	-0.212	0.0060	0.734(a)

Fuente: Sistema de Predicción de Propiedades, SPP 2.0

En esta matriz se tiene en cuenta la diagonal principal, si todos los elementos son mayores o iguales a 0.30 se consideran dentro del análisis, si algún dato es menor a 0.30 debe eliminarse y realizar de nuevo el análisis de componentes principales. En este caso S/C es de 0.299 de acuerdo con lo dicho anteriormente tendría que eliminarse, pero como el KMO es mayor de 0.7 y S/C es una variable importante para el usuario se considerará en el análisis.

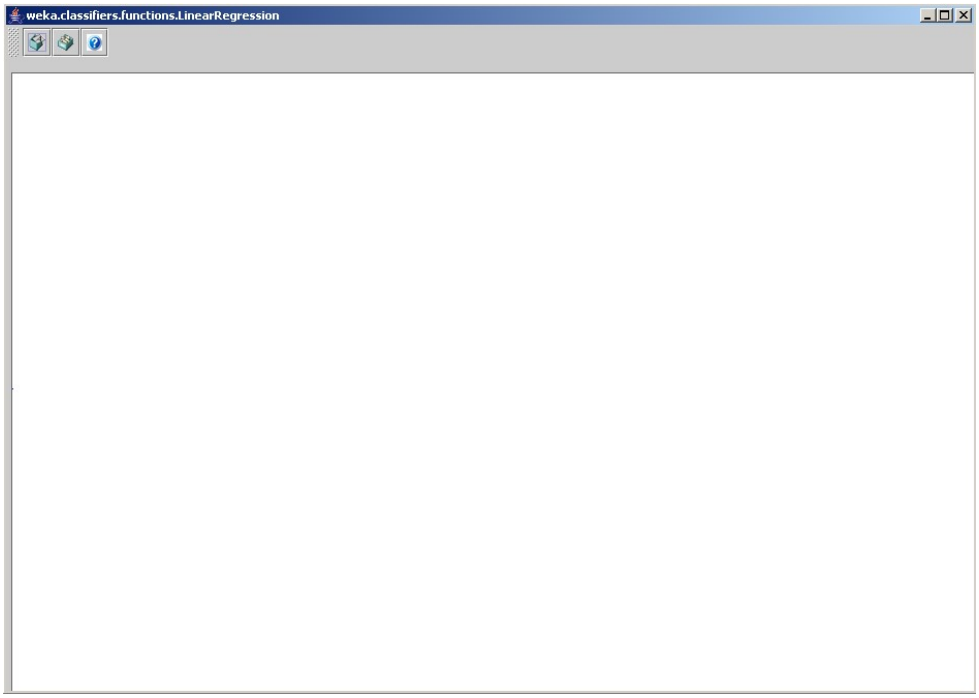
Las líneas de código para la clase Componentes en la que se calcula, el test de esfericidad, el índice de KMO, el determinante de la matriz de correlación y la matriz de correlaciones anti-imagen.

```
/*
 * Componentes.java
 *
 * Created on 25 de septiembre de 2006, 10:31 PM
 *
 * To change this template, choose Tools | Template Manager
 * and open the template in the editor.
 */
package InterfazWeka;

public class Componentes {

    public Componentes() {
    }
    public static void main(String[] args) {
    }
    public static double determinante(double[] matriz) {
        int r = (int)Math.sqrt(matriz.length);
        double[][] x = new double[r][r];
        for (int k = 0; k < r; k++) {
            for (int i = 0; i < r; i++) {
                x[k][i] = matriz[k + (i * r)];
            }
        }
        for (int k = 0; k < x.length - 1; k++) {
            for (int i = k + 1; i < x[0].length; i++) {
                for (int j = k + 1; j < x.length; j++) {
                    x[i][j] -= x[i][k] * x[k][j] / x[k][k];
                }
            }
        }
        double deter = 1.0;
        for (int i = 0; i < x.length; i++){
            deter *= x[i][i];
        }
        return deter;
    }
    public static double bartlett(int n, double[] matriz) {
        int r = (int)Math.sqrt(matriz.length);
        double det = determinante(matriz);
        return - (n - 1 - (((2 * r) + 5) / 6)) * Math.log(det);
    }
    public static double kmo(double[] antiImagen, double[] correlacion)
    {
        int r = (int)Math.sqrt(antiImagen.length);
        double rij = 0;
        double aij = 0;
        for (int j = 0; j < r; ++j) {
            for (int i = 0; i < r; ++i) {
                if (i != j) {
                    rij += correlacion[i + (j * r)] * correlacion[i + (j * r)];
                    aij += antiImagen[i + (j * r)] * antiImagen[i + (j * r)];
                }
            }
        }
        return rij / (rij + aij);
    }
 public static double[] calcular(double[] antiImagen, double[] correlacion) {
        int r = (int)Math.sqrt(correlacion.length);
        int[] coords = new int[r];
        for (int j = 0; j < r; ++j) {
```

```
            for (int i = j + 1; i < r; ++i) {
                coords[0] = i;
                coords[1] = j;
                for (int k = 2, l = 0; k < coords.length; ++k, ++l) {
                    while (l == i || l == j)
                        ++l;
                    coords[k] = l;
                }
                antiImagen[i + (j * r)] = secundariosAntiImagen(coords, correlacion, r) * -1;
                antiImagen[j + (i * r)] = antiImagen[i + (j * r)];
            }
        }
        for (int j = 0; j < r; ++j) {
            double rij = 0;
            double aij = 0;
            for (int i = 0; i < r; ++i) {
                if (i != j) {
                    rij += correlacion[i + (j * r)] * correlacion[i + (j * r)];
                    aij += antiImagen[i + (j * r)] * antiImagen[i + (j * r)];
                }
            }
            antiImagen[j + (j * r)] = rij / (rij + aij);
        }
        return antiImagen;
    }
    public static double secundariosAntiImagen(int[] coords, double[] matriz, int r) {
        double r12, r13, r23;
       if (coords.length > 3) {
            int[] ncoords = new int[coords.length - 1];

            for (int i = 0; i < coords.length - 1; ++i)
                ncoords[i] = coords[i];
            r12 = secundariosAntiImagen(ncoords, matriz, r);
            ncoords[0] = coords[0];
            ncoords[1] = coords[coords.length - 1];
            r13 = secundariosAntiImagen(ncoords, matriz, r);
            ncoords[0] = coords[1];
            ncoords[1] = coords[coords.length - 1];
            r23 = secundariosAntiImagen(ncoords, matriz, r);
        }
        else {
            int i = coords[0];
            int j = coords[1];
            int k = coords[2];
            r12 = matriz[i + (j * r)];
            r13 = matriz[i + (k * r)];
            r23 = matriz[j + (k * r)];
        }
        return (r12 - (r13 * r23)) /
            ( Math.sqrt(1 - (r13 * r13)) * Math.sqrt(1 - (r23 * r23) ) );
    }
}
```

La interfaz de respuesta es controlada por la clase Fromrespuesta. En los que se pueden ver los resultados del prototipo de cada una de las técnicas descritas anteriormente es la siguiente (ver Figura 26).

Figura 26. Interfaz de Respuesta



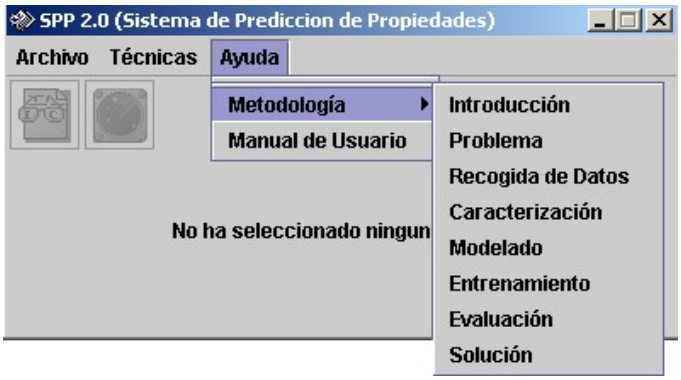
Fuente: Sistema de Predicción de Propiedades, SPP 2.0

El primer botón que se ve en la Figura 26 de la interfaz de respuesta es: “guardar los resultados actuales”, esta acción pide al usuario indicar la ruta y el nombre del archivo .rtf para guardar los resultados que se presentan.

El segundo botón de la interfaz de respuesta es: “añadir los resultados actuales”, esta acción guarda en una dirección por *default* (SPP/temp/resultadosglobales) los resultados actuales.

El tercer botón de la interfaz de respuesta es: “interpretar los resultados”, esta acción muestra cómo el usuario debe interpretar los resultados de cada una de las técnicas. Este botón fue diseñado por los autores y usa la clase VentanaAyuda.

Figura 27. Menú Ayuda



Fuente: Sistema de Predicción de Propiedades, SPP 2.0

El menú que se muestra de la Figura 27 es un componente totalmente desarrollado por los autores así como la descripción de cada uno de ellos, que se detallará en el ítem 7: Propuesta de una serie de pasos. El manual de usuario puede verse en el anexo G. Para esto se usa la clase VentanaAyuda:

```
/*
* VentanaAyuda.java
*
*/

package Ayuda;
import javax.swing.JFrame;
import javax.swing.*;
import javax.swing.event.*;
import javax.swing.text.html.*;
import java.net.*;
import javax.swing.JScrollPane;

public class VentanaAyuda extends JFrame {
private JEditorPane jEditorPane = null;
private String path=null;
private String titulo=null;
private JScrollPane jScrollPane = null;
public VentanaAyuda(String path,String titulo){
super();
this.path=path;
this.titulo=titulo;
initialize();
}
private void initialize() {
this.setBounds(100,30,800,600);
this.setContentPane(getJScrollPane());
this.setVisible(true);
this.setTitle(titulo);
}
private JEditorPane getJEditorPane() {
```

```
if (jEditorPane == null) {
URL url=null;
try {
url=getClass().getResource(path);
jEditorPane = new JEditorPane(url);
jEditorPane.setFont(new java.awt.Font("Verdana", 0, 11));
jEditorPane.setCaretPosition(0);
jEditorPane.setAlignmentX(1.0F);
jEditorPane.setEditable(false);
jEditorPane.addHyperlinkListener(new HtmlEventos());
} catch (Exception e) {
JOptionPane.showMessageDialog(null,"Error al cargar la página");
}

}
return jEditorPane;
}
private JScrollPane getJScrollPane() {
if (jScrollPane == null) {
jScrollPane = new JScrollPane(getJEditorPane());
}
return jScrollPane;
}
public static void main(String args[]){
}
}//  @jve:decl-index=0:visual-constraint="10,10"//end class
class HtmlEventos implements HyperlinkListener {
public void hyperlinkUpdate(HyperlinkEvent e) {
if (e.getEventType() == HyperlinkEvent.EventType.ACTIVATED) {
JEditorPane pane = (JEditorPane) e.getSource();
if (e instanceof HTMLFrameHyperlinkEvent) {
HTMLFrameHyperlinkEvent  evt = (HTMLFrameHyperlinkEvent)e;
HTMLDocument doc = (HTMLDocument)pane.getDocument();
doc.processHTMLFrameHyperlinkEvent(evt);
} else {
try {
pane.setFont(new java.awt.Font("verdana",0,11));
pane.setPage(e.getURL());
} catch (Throwable t) {
t.printStackTrace();
}
}
}
}
}
```

– **Descripción Técnica**: se hará una descripción a nivel de ingeniería de software con la notación de UML[29] donde se representa la descripción general ya mencionada en casos de uso, diagramas de secuencia y diagramas de clase.

[29] RUMBAUGHT, James; JACOBSON, Ivar y BOOCH, Grady. El Lenguaje Unificado de Modelado. Manual de Referencia. Edición año 2000. Editorial Addison Wesdey.

- **Descripción del usuario**

Representante	Martha Parra
Descripción	Ingeniera química, su función está relacionada con el análisis de grandes volúmenes de datos provenientes empresas petrolíferas o de laboratorios como la planta DEMEX del ICP, donde está más al tanto, de los factores o materiales influyentes de los hidrocarburos.
Responsabilidades	Realizar una interpretación positiva de los datos arrojados por el prototipo, para que el análisis realizado por éste sea de utilidad.

Fuente: autores del proyecto

- **Caso de uso**: el diagrama de casos de uso representa cómo el usuario opera o interactúa sobre el sistema en cuestión. El diagrama de caso de uso y la descripción de cada uno de los casos de uso se encuentran en el Anexo A.

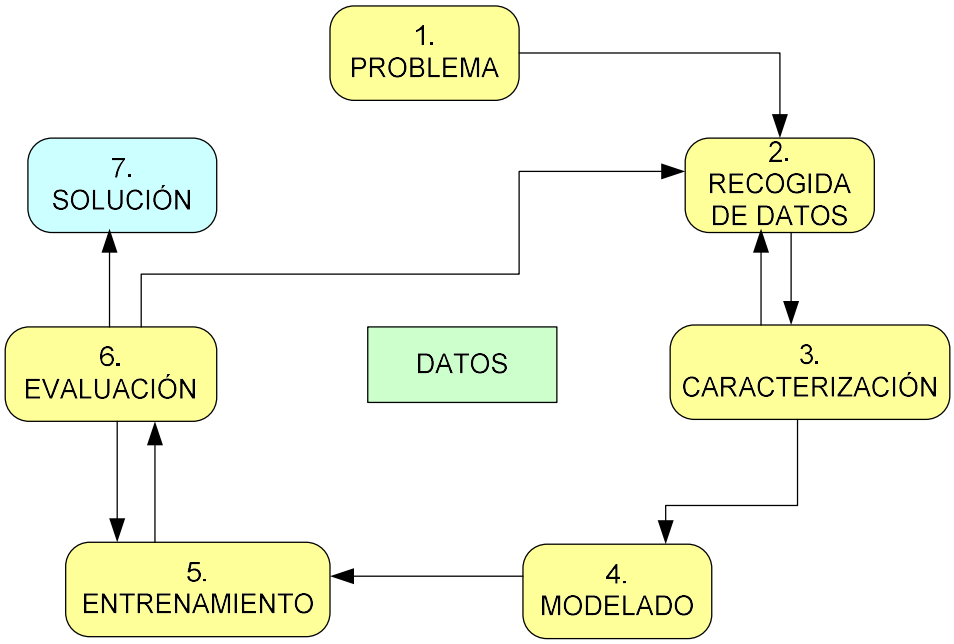
- **Diagrama de secuencia**: estos representan la forma como los actores y las clases se comunican entre sí ante la petición de un evento, en esto interviene toda una secuencia de llamadas de donde se obtienen unas responsabilidades. Estos diagramas están compuestos por objetos o actores, mensajes entre objetos y mensajes de un objeto a sí mismo. Los diagramas en el Anexo B fueron hechos por el autor y se usaron los nombres de los métodos de cada una de las clases. Las clases que están con color gris indican que han sido modificadas agregándoles nuevas líneas de código y las de rosado son las clases creadas por el usuario.

- **Diagrama de clase**: los diagramas de clases ayudan a visualizar las clases que están interactuando en un sistema, como se muestra en el Anexo C.

3. PROPUESTA DE UNA SERIE DE PASOS

Con base en la metodología CRISP-DM y las lecturas hechas sobre análisis multivariante se decide implementar los pasos planteados en la Figura 28, los cuales guiarán al usuario en la aplicación de las Técnicas de Minería y podrán ser adaptados a las necesidades de cada organización.

Figura 28. Modelo de pasos



Fuente: autores del proyecto

Como se puede observar, los datos son el centro de este proceso ya que son la unidad fundamental en la aplicación de Minería de Datos. Las flechas indican las secuencias a seguir, en algunas etapas existe una continua iteración que permitirá encontrar la mejor solución. La intervención del usuario es primordial ya que él será el encargado de interpretar los resultados en cada una de las etapas, así como de analizar, concluir y hallar una solución con éxito.

Se espera que estos pasos planteados faciliten o ayuden al usuario en la aplicación de técnicas de Minería de Datos así como la compresión de los resultados.

3.1 PROBLEMA

En esta etapa se recomienda plantear el problema, los requerimientos, los objetivos, el presupuesto (tiempo, personal), la terminología y los beneficios de realizar el proyecto. Básicamente el problema consiste en la implementación de una serie de pasos para la aplicación de técnicas de Minería de Datos en el análisis de información generada por la planta Demex de ECOPETROL, así como determinar cuáles variables predicen mejores rendimientos para el DMO (residuo de un proceso en el tratamiento del crudo).

3.2 RECOGIDA DE DATOS

Esta fase inicia con una colección de datos y procede con actividades para familiarizar con los datos, identificar problemas de calidad y descubrir las primeras potencialidades en los datos o detectar subconjuntos. Se deben recopilar los datos, descripción de los mismos, exploración y verificación de calidad.

El usuario debe extraer de los datos que se van a usar en todo el proceso, esto se denomina *datawarehouse*[30] (DW). El DW puede verse como una bodega donde están almacenados todos los datos necesarios para realizar las funciones de gestión de la empresa, de manera que puedan utilizarse fácilmente según se necesiten. Esta extracción puede ser apoyada por el ingeniero de sistemas de la organización o aquel que esté encargado del mantenimiento de la base de datos.

El contenido de los datos, la organización y estructura son dirigidos a satisfacer las necesidades de información de analistas. El objetivo del DW será el de satisfacer los requerimientos de información interna de la empresa para una mejor gestión. El contenido de los datos, la organización y estructura son dirigidos a satisfacer las necesidades de información de los analistas.

El *datawarehouse* debe estar completo, es decir, no debe haber ningún campo en blanco si existe esto el usuario esta en la capacidad de llenar estos campos a través de patrones de coincidencia. Las variables extraídas desde el silab por el usuario están en la Tabla 7.

[30] WOLFF Carmen Gloria. La Tecnología Datawarehousing. 1999 [online, Articulo]. [Citado el 27 de agosto 2006]. Disponible en Internet: <http://www.inf.udec.cl/revista/ediciones/edicion3/cwolff.PDF> p. 2.

Tabla 7. Variables generales de uso

NOMBRE	ALIAS	CLASIFICACIÓN
Solvente/Carga	S/C	Independiente
Solvente	C3	Independiente
Solvente	i-C4	Independiente
Solvente	n-C4	Independiente
Temperatura	Trect	Independiente
Densidad a 15 Gr C	D-carga	Independiente
Residuo carbón micro	CCR-carga	Independiente
Níquel	Ni-carga	Independiente
Vanadio	V-carga	Independiente
Vanadio a 50	V50-carga	Independiente
Carbón aromático	CarAro-carga	Independiente
Penetración a 25 Gr C	Penet-carga	Independiente
5%	PIE-carga	Independiente
Nitrógeno básico en hidrocarburos	N2b-carga	Independiente
Saturados	SatCarga	Independiente
Aromáticos	AromCarga	Independiente
Resinas	ResCarga	Independiente
Asfáltenos	AsfCarga	Independiente
Hierro	FE-carga	Independiente
Azufre	S-carga	Independiente
10%	T10	Independiente
30%	T30	Independiente
50%	T50	Independiente
	BCMI-carga	Independiente
Rendimiento del DMO	RendvDMO	Dependiente
Categorización del DMO	CatRendDMO	Dependiente

Fuente: autores del proyecto

3.3 CARACTERIZACIÓN

Seleccionar el conjunto de rasgos y valores que caracterizan los datos según su integración, selección, limpieza y transformación. Se debe tener en cuenta la normalidad, homocedasticidad (se refiere al supuesto de que las variables dependientes exhiban iguales niveles de varianza a lo largo del rango del predictor de las variables) y linealidad (la ausencia de linealidad provoca que el coeficiente de correlación no mida adecuadamente la relación entre los pares de variables) de los datos. Los datos puede ser clasificados en:

Escalas nominales o categóricas. En este caso, los números se comportan como etiquetas. En esta escala la variable CatRendDMO ya que su valor está dado por (bajo, medio o alto).

Escalas ordinales o numéricas. No sólo consiguen distinguir entre valores, sino que además establece un orden entre ellas. Clasifican las variables independientes que tengan valor numérico así como la dependiente RendvDMO.

Estas clasificaciones se pueden agrupar de la siguiente manera **escalas métricas** (escalas de intervalo y razón)**, escalas no métricas** (escalas nominales y ordinales). La caracterización de las variables puede verse en la Tabla 8.

Tabla 8. Caracterización de las variables

ALIAS	CLASIFICACIÓN	CARACTERÍSTICA
S/C	Independiente	Numérica
C3	Independiente	Numérica
i-C4	Independiente	Numérica
n-C4	Independiente	Numérica
Trect	Independiente	Numérica
D-carga	Independiente	Numérica
CCR-carga	Independiente	Numérica
Ni-carga	Independiente	Numérica
V-carga	Independiente	Numérica
V50-carga	Independiente	Numérica
CarAro-carga	Independiente	Numérica
Penet-carga	Independiente	Numérica
PIE-carga	Independiente	Numérica
N2b-carga	Independiente	Numérica
SatCarga	Independiente	Numérica
AromCarga	Independiente	Numérica
ResCarga	Independiente	Numérica
AsfCarga	Independiente	Numérica
FE-carga	Independiente	Numérica
S-carga	Independiente	Numérica
T10	Independiente	Numérica
T30	Independiente	Numérica
T50	Independiente	Numérica
BCMI-carga	Independiente	Numérica
RendvDMO	Dependiente	Numérica
CatRendDMO	Dependiente	Nominal

Fuente: autores del proyecto

3.4 MODELADO

Se debe combinar la información de las características de los datos. No hay modelos mejores que otros; de acuerdo a las características del problema y de los datos de búsqueda en el modelo, es un proceso de ensayo y error.

En el objeto de estudio “Análisis de información generada por la planta Demex de ECOPETROL” y en la aplicación de las técnicas de análisis multivariado y uso del prototipo se consideran dos modelos en orden jerárquico que pueden verse el en la Figura 29 y Figura 30 respectivamente.

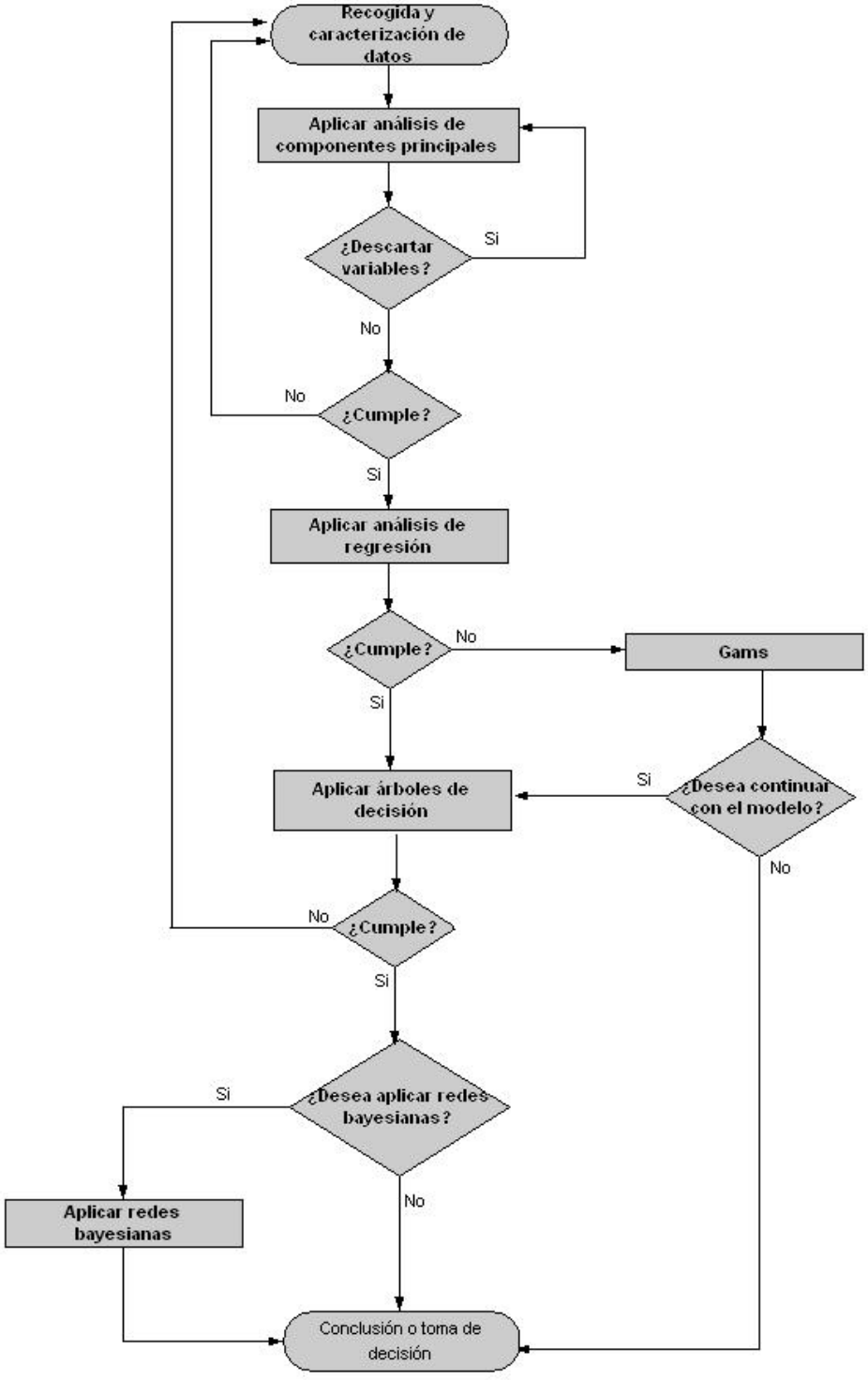
Iniciando con la extracción de datos desde la base de datos *Silab* con las variables independientes (S/C, C3, i-C4, n-C4,Trect, D-carga, CCR-carga, Ni-carga, V-carga, V50-carga, CarAro-carga, Penet-carga, PIE-carga, N2b-carga, SatCarga, AromCarga, ResCarga, AsfCarga, FE-carga, S-carga, T10, T30, T50, BCMI-carga), y dependientes (RendvDMO, CatRendDMO).

De las variables dependientes se busca predecir el mejor rendimiento, a partir de los atributos de las variables independientes a través de los modelos que se presentan a continuación y su respectiva evaluación.

La primera técnica a aplicar es el Análisis de Componentes Principales, porque ayuda a reducir variables; sólo se podrá continuar cuando se cumplan todos los criterios de la técnicas, si esto no ocurre se debe tomar otra base de datos. La segunda técnica es la de Análisis de Regresión para conocer el valor de la variable dependiente con respecto a las independientes. Tercero, el Árbol de Decisión para saber en que condiciones el rendimiento del DMO es alto, y por último Redes Bayesianas para conocer la probabilidad de las variables independientes sobre la dependiente. El proceso *Gams* se llevará a cabo si no se consigue un modelo lineal óptimo con el Análisis de Regresión. (Ver figura 29)

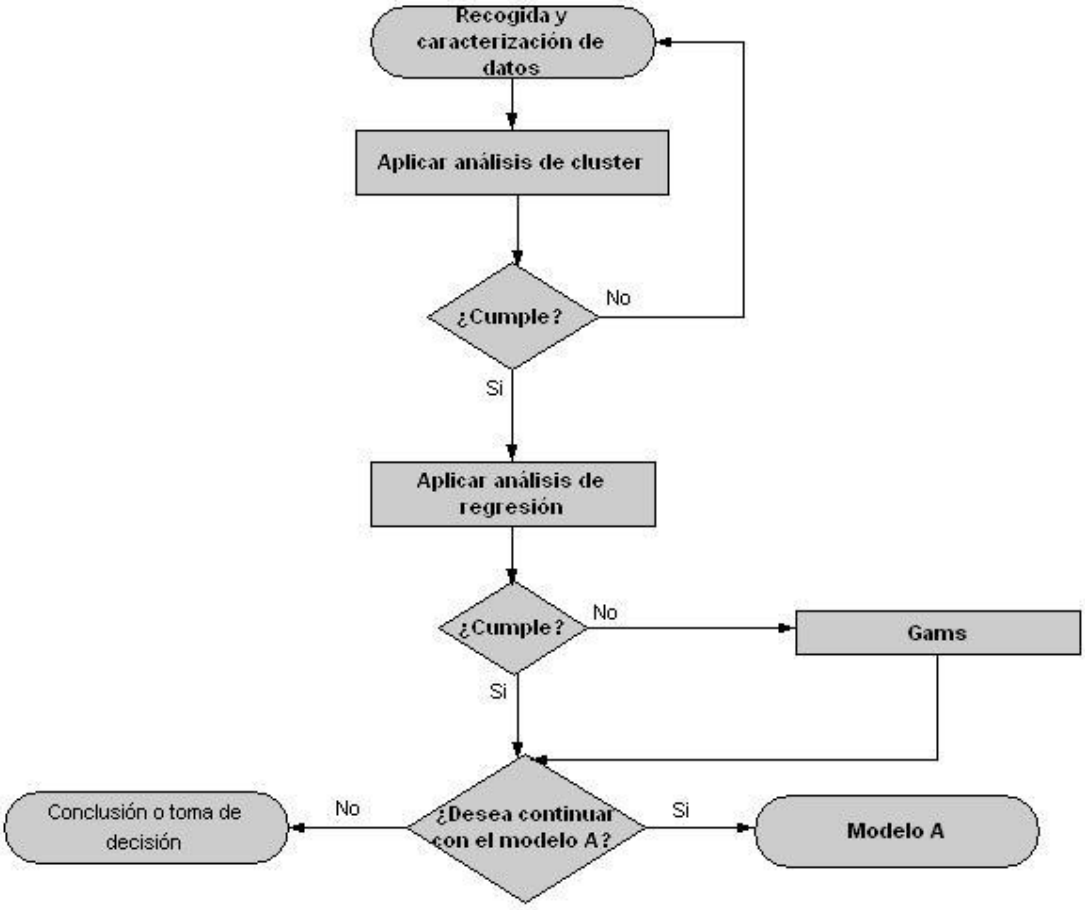
El modelo de la figura 30 indica que se deben agrupar las variables por su homogeneidad dentro del grupo, y por heterogeneidad con los demás *cluster*. Después aplicar un Análisis de Regresión, el usuario puede concluir o continuar con el *modelo A* de la figura 29. El proceso *Gams* se usará cuando no se obtenga un modelo lineal adecuado en el análisis de regresión.

Figura 29. Modelo A



Fuente: autores del proyecto

Figura 30. Modelo B



Fuente: autores del proyecto

3.5 ENTRENAMIENTO

Esta etapa corresponde a la explicación de los resultados de cada una de las técnicas a aplicar, para que el usuario pueda interpretarlos cuando esté en la etapa 6, Evaluación.

3.5.1 Resultados de la técnica análisis de regresión el Análisis de Regresión es una técnica estadística utilizada para estudiar la relación entre una sola variable dependiente y varias independientes. El objetivo de esta técnica es usar las variables independientes, cuyos valores se conocen, para predecir el de la variable dependiente. Los resultados de esta técnica constan del modelo lineal,

así como la aplicación de este con el primer registro del *dataset* y el sumario donde se muestran los errores. Esto se puede apreciar en la Figura 31.

Las variables que se usaron para explicar los resultados son; S/C, C3, Trect, D-carga, CCR-carga, Ni-carga, V-carga, V50-carga, carAro-carga y penet-carga. Se muestra sólo el texto de los resultados sin la interfaz para una mejor claridad en la presentación y entendimiento de los resultados arrojados por el prototipo.

Figura 31. Resultados de la técnica análisis de regresión

Linear Regression Model

RendvDMO = 1.8857 * S/C - 0.8558 * C3 - 0.3577 * Trect - 490.7619 * D-carga - 1.0295 * CCR-carga - 0.0613 * Ni-carga - 0.0276 * V-carga + 0.8391 * V50-carga + 1.4144 * CarAro-carga + 540.4457

Al evaluar el modelo remplazando las variables por los valores de la primera línea del *Dataset* se encontró que:

RendvDMO = 73.73351926129703

=== Summary ===

Correlation coefficient	0.8694
Mean absolute error	2.7951
Root mean squared error	4.3773
Relative absolute error	40.4378 %
Root relative squared error	49.4081 %
Total Number of Instances	202

Fuente: Sistema de Predicción de Propiedades, SPP 2.0

En el modelo lineal se ve claramente que la variable rendVDMO depende de la variable D-carga y parcialmente por las otras variables. En el *summary* se pueden apreciar el coeficiente de correlación (*Correlation coefficient*) de 0.8694, este valor es muy bueno ya que el coeficiente de relación indica la fuerza de la relación entre las variables independientes y la variable dependiente. Puede tomar valores entre -1 y +1: con +1 indica una relación positiva perfecta, 0 ausencia de relación y -1 indica una relación inversa perfecta.

La medida del error absoluto (*Mean absolute error*) es el promedio de la magnitud de los errores individuales sin tomar en cuenta el signo. O es igual a la imprecisión que acompaña a la medida.

Root mean squared error da la medida de las diferencias en promedio de los valores pronosticados y los observados.

3.5.2 Resultados de la técnica análisis de componentes principales es una técnica de reducción de datos, es decir, pretende pasar de ese número elevado de variables, a un número más pequeño de elementos explicativos, factores que le permitan explicar de una manera más sencilla esa realidad. Es decir, ante un banco de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible. Esta técnica es adecuada cuando se trata de resumir la mayor parte posible de la información inicial (varianza) en el menor número de factores posibles.

Esta técnica arroja en los resultados los elementos: Matriz de correlación, eigenvalores, eigenvectores, determinante de la matriz de correlaciones, coeficiente de esfericidad de Bartlett, KMO, y matriz anti-imagen de correlaciones; estos elementos se pueden observar en la Figura 32, y a continuación su respectiva explicación.

Figura 32. Resultados de la técnica de análisis de componentes principales

Principal Components Attribute Transformer

Correlation matrix

1	0.07	0.03	0	0.02	-0.01	0	0.02	-0.01	-0.01
0.07	1	-0.09	0.07	0.1	0.1	-0.11	-0.03	0.07	-0.09
0.03	-0.09	1	0.06	0.02	0.05	0.07	-0.01	0.09	0
0	0.07	0.06	1	0.88	0.68	0.63	0.55	0.82	-0.71
0.02	0.1	0.02	**0.88**	1	0.59	0.52	0.57	0.59	-0.79
-0.01	0.1	0.05	**0.68**	**0.59**	1	0.6	0.43	0.56	-0.67
0	-0.11	0.07	**0.63**	**0.52**	0.6	1	0.44	0.4	-0.55
0.02	-0.03	-0.01	**0.55**	**0.57**	**0.43**	**0.44**	1	0.05	-0.59
-0.01	0.07	0.09	**0.82**	**0.59**	**0.56**	**0.4**	0.05	1	-0.38
-0.01	-0.09	0	-0.71	**-0.79**	**-0.67**	**-0.55**	**-0.59**	**-0.38**	1

Sugerencia:

Se recomienda trabajar con las variables de los componentes cuyos eigenvalues sean superiores a la unidad.

Eigenvalue proportion cumulative

4.50638	0.45064	0.45064	-0.004S/C-0.035C3-0.029Trect-0.446D-carga-0.422CCR-carga-0.382Ni-carga-0.345V-carga-0.3V50-carga-0.326CarAro-carga+0.4 Penet-carga
1.14279	0.11428	0.56492	-0.274S/C-0.79C3+0.469Trect-0.025D-carga-0.078CCR-carga-0.023Ni-carga+0.239V-carga+0.095V50-carga-0.069CarAro-carga+0.051Penet-carga
1.09513	0.10951	0.67443	-0.126S/C-0.153C3-0.541Trect-0.137D-carga+0.029CCR-carga-0.059Ni-carga+0.07 V-carga+0.546V50-carga-0.547CarAro-carga-0.207Penet-carga
1.01834	0.10183	0.77626	0.872S/C-0.039C3+0.336Trect-0.066D-carga+0.004CCR-carga-0.05Ni-carga+0.055V-carga+0.231V50-carga-0.239CarAro-carga-0.062Penet-carga
0.82996	0.083	0.85926	-0.382S/C+0.503C3+0.611Trect-0.097D-carga+0.03 CCR-carga+0.012Ni-carga-0.142V-carga+0.28 V50-carga-0.313CarAro-carga-0.149Penet-carga
0.5621	0.05621	0.91547	0.001S/C+0.202C3-0.031Trect-0.251D-carga-0.402CCR-carga+0.489Ni-carga+0.646V-carga-0.212V50-carga-0.178CarAro-carga+0.031Penet-carga
0.38604	0.0386	0.95407	-0.045S/C+0.235C3-0.002Trect+0.209D-carga+0.038CCR-carga-0.594Ni-carga+0.553V-carga+0.215V50-carga+0.082CarAro-carga+0.431Penet-carga

Eigenvectors

V1	V2	V3	V4	V5	V6	V7	
-0.0036	-0.274	-0.1259	0.8724	-0.3816	0.0008	-0.0451	S/C
-0.0346	-0.7902	-0.1525	-0.0387	0.5025	0.2021	0.2354	C3
-0.0288	0.4687	-0.5405	0.3364	0.6105	-0.031	-0.0022	Trect
-0.4464	-0.0254	-0.1373	-0.0663	-0.0968	-0.2505	0.2088	D-carga
-0.4224	-0.0781	0.0294	0.0045	0.0302	-0.4018	0.0382	CCR-carga
-0.3817	-0.0234	-0.0589	-0.0498	0.0116	0.4889	-0.5944	Ni-carga
-0.3448	0.2391	0.0702	0.0549	-0.142	0.6461	0.553	V-carga
-0.2997	0.0949	0.5461	0.2308	0.2801	-0.2119	0.2146	V50-carga
-0.3256	-0.0689	-0.5471	-0.2391	-0.3131	-0.1777	0.0825	CarAro-carga
0.3999	0.0511	-0.2073	-0.0625	-0.1485	0.0314	0.4307	Penet-carga

EL DETERMINANTE DE LA MATRIZ ES:

4.6303301576024376E-4

COEFICIENTE DE ESFERICIDAD DE BARTLETT

1535.5424396165797

KMO

0.585365511948193

MATRIZ ANTI-IMAGEN DE CORRELACIONES

S/C	C3	Trect	D-carga	CCR-carga	Ni-carga	V-carga	V50-carga	CarAro-carga	Penet-carga
0.204(a)	-0.085	-0.039	0.063	-0.057	0.042	-0.046	-0.06	-0.055	0.0060
-0.085	0.224(a)	0.085	-0.171	0.054	-0.18	0.269	0.196	0.173	-0.01
-0.039	0.085	0.719(a)	-0.0010	0.011	-0.016	-0.029	0.0040	-0.018	-0.024
0.063	-0.171	-0.0010	0.549(a)	-0.699	0.295	-0.553	-0.84	-0.942	0.213
-0.057	0.054	0.011	-0.699	0.7(a)	-0.057	0.371	0.424	0.516	0.268
0.042	-0.18	-0.016	0.295	-0.057	0.731(a)	-0.391	-0.355	-0.42	0.446
-0.046	0.269	-0.029	-0.553	0.371	-0.391	0.612(a)	0.396	0.492	-0.029
-0.06	0.196	0.0040	-0.84	0.424	-0.355	0.396	0.403(a)	0.878	-0.148
-0.055	0.173	-0.018	-0.942	0.516	-0.42	0.492	0.878	0.401(a)	-0.294
0.0060	-0.01	-0.024	0.213	0.268	0.446	-0.029	-0.148	-0.294	0.847(a)

Fuente: Sistema de Predicción de Propiedades, SPP2.0

Matriz de correlaciones: Es la matriz que indica las correlaciones entre las variables, por ejemplo hay correlación entre dos variables cuando éstas cambian de tal modo que los valores que toma una de ellas son, hasta cierto punto,

predecibles a partir de los que toma la otra. Los números sombreados expresan correlación entre las variables. Se debe observar por debajo o por encima de la diagonal principal el numero de ellas, si existen la mitad + 1 de ellas mayores de 0.3, se dice que son significativas lo que permite profundizar en el análisis. En el ejemplo hay 19 de 35.

Correlation matriz

1	0.07	0.03	0	0.02	-0.01	0	0.02	-0.01	-0.01	
0.07	1	-0.09	0.07	0.1	0.1	-0.11	-0.03	0.07	-0.09	
0.03	-0.09	1	0.06	0.02	0.05	0.07	-0.01	0.09	0	
0	0.07	0.06	1	0.88	0.68	0.63	0.55	0.82	-0.71	
0.02	0.1	0.02	**0.88**	1	0.59	0.52	0.57	0.59	-0.79	
-0.01	0.1	0.05	**0.68**	**0.59**	1	0.6	0.43	0.56	-0.67	
0	-0.11	0.07	**0.63**	**0.52**	**0.6**	1	0.44	0.4	-0.55	
0.02	-0.03	-0.01	**0.55**	**0.57**	**0.43**	**0.44**	1	0.05	-0.59	
-0.01	0.07	0.09	**0.82**	**0.59**	**0.56**	**0.4**	0.05	1	-0.38	
-0.01	-0.09	0	**-0.71**	**-0.79**	**-0.67**	**-0.55**	**-0.59**	**-0.38**	1	

Eigenvalores: Son los valores explicados de los factores. Estos determinan cuales factores se deben tomar para el análisis de componentes, dependiendo de la varianza acumulada, se debe tomar en consideración los que sean mayores que 1 para el ejemplo se toman los primeros 4, y con estos eigenvalores se explica el 77% de la varianza.

Eigenvalue	proportion	cumulative
4.50638	0.45064	0.45064
1.14279	0.11428	0.56492
1.01834	0.10183	0.77626
1.01834	0.10183	**0.77626**

Eigenvector: Los valores que conforman el *eigenvector* explican a la variable en cada uno de ellos; para el ejemplo anterior los eigenvectores que se tienen en cuenta son los cuatro primeros, y estos contienen una lista de valores para cada variable, donde el valor más alto de la variable significa que esta mejor representada en el eigenvector correspondiente, por ejemplo que la variable S/C esta explicada en v4, porque su valor corresponde a 0.8724, mientras que en el v1 corresponde a 0.0036, en el v2 0.274, y en el v3 0.1259, como se indica a continuación:

Eigenvectors

V1	V2	V3	**V4**	
-0.0036	-0.274	-0.1259	**0.8724**	S/C
-0.0346	**-0.7902**	-0.1525	-0.0387	C3
-0.0288	0.4687	**-0.5405**	0.3364	Trect
-0.4464	-0.0254	-0.1373	-0.0663	D-carga
-0.4224	-0.0781	0.0294	0.0045	CCR-carga
-0.3817	-0.0234	-0.0589	-0.0498	Ni-carga
-0.3448	0.2391	0.0702	0.0549	V-carga
-0.2997	0.0949	**0.5461**	0.2308	V50-carga
-0.3256	-0.0689	**-0.5471**	-0.2391	CarAro-carga
0.3999	0.0511	-0.2073	-0.0625	Penet-carga

Teniendo en cuenta esta clasificación, V1 está conformado por: D-carga, CCR-carga, Ni-carga, V-carga, penet-carga V2 por C3, V3 por Trect, V50-carga, CarAro-carga y como se mencionó anteriormente, V4 por S/C

El determinante de la matriz debe acercarse a cero porque indica altas inter-correlaciones, pero no debe ser cero. En el ejemplo se acerca a cero y se acepta el análisis.

El índice de Kayes Meyer Olkin (KMO) debe ser mayor a 0.7 para que la aceptación sea conveniente y exprese alta correlación entre las variables; si es mayor a 0.7 el grado de aceptación es superior, menor de 0.5 no es factible el análisis y entre 0.5 y 0.7 la aceptación es media. En el ejemplo se tiene una aceptación media.

KMO
0.585365511948193

En la matriz de correlación parcial anti-imagen se tiene en cuenta la diagonal principal porque contiene las MSA (Medida de Suficiencia de Muestreo) para cada variable. La experiencia en este análisis indica que el MSA debe ser mayor de 0.3 para aceptar la variable. Y los valores fuera de la diagonal son correlaciones parciales entre las variables. De esta manera se deben eliminar la variable C3.

MATRIZ ANTI-IMAGEN DE CORRELACIONES

S/C	C3	Trect	D-carga	CCR-carga	Ni-carga	V-carga	V50-carga	CarAro-carga	Penet-carga

0.204(a)	-0.085	-0.039	0.063	-0.057	0.042	-0.046	-0.06	-0.055	0.0060
-0.085	0.224(a)	0.085	-0.171	0.054	-0.18	0.269	0.196	0.173	-0.01
-0.039	0.085	0.719(a)	-0.0010	0.011	-0.016	-0.029	0.0040	-0.018	-0.024
0.063	-0.171	-0.0010	0.549(a)	-0.699	0.295	-0.553	-0.84	-0.942	0.213
-0.057	0.054	0.011	-0.699	0.7(a)	-0.057	0.371	0.424	0.516	0.268
0.042	-0.18	-0.016	0.295	-0.057	0.731(a)	-0.391	-0.355	-0.42	0.446
-0.046	0.269	-0.029	-0.553	0.371	-0.391	0.612(a)	0.396	0.492	-0.029
-0.06	0.196	0.0040	-0.84	0.424	-0.355	0.396	0.403(a)	0.878	-0.148
-0.055	0.173	-0.018	-0.942	0.516	-0.42	0.492	0.878	0.401(a)	-0.294
0.0060	-0.01	-0.024	0.213	0.268	0.446	-0.029	-0.148	-0.294	0.847(a)

Eliminando la variable C3, se puede observar que el KMO sube, se toman 3 eigenvectores, explicando el 73% de la varianza acumulada, V1 está dado por D-carga, CCR-carga, Ni-carga, V-carga y penet-carga V2 por V50-carga y CarAro-carga y V3 por S/C; y los MSA (diagonal principal de la matriz anti-imagen) se mantienen o aumentan. No se elimina S/C por ser una variable de suma importancia para el usuario. Si el usuario determina conveniente tomar otra muestra y hacer de nuevo el análisis podrá hacerlo sin ningún problema. Los datos que estén sombreados y de otros colores son usados para explicar, es decir, no son parte de los resultados arrojados por el prototipo.

Principal Components Attribute Transformer

Correlation matrix

1	0.03	0	0.02	-0.01	0	0.02	-0.01	-0.01
0.03	1	0.06	0.02	0.05	0.07	-0.01	0.09	0
0	0.06	1	0.88	0.68	0.63	0.55	0.82	-0.71
0.02	0.02	**0.88**	1	0.59	0.52	0.57	0.59	-0.79
-0.01	0.05	**0.68**	**0.59**	1	0.6	0.43	0.56	-0.67
0	0.07	**0.63**	**0.52**	**0.6**	1	0.44	0.4	-0.55
0.02	-0.01	**0.55**	**0.57**	**0.43**	**0.44**	1	0.05	-0.59
-0.01	0.09	**0.82**	**0.59**	**0.56**	**0.4**	0.05	1	-0.38
-0.01	0	**-0.71**	**-0.79**	**-0.67**	**-0.55**	**-0.59**	**-0.38**	1

Sugerencia:

Se recomienda trabajar con las variables de los componentes cuyos eigenvalues sean superiores a la unidad.

eigenvalue	proportion	cumulative	
4.5022	0.50024	0.50024	-0.003S/C-0.03Trect-0.446D-carga-0.422CCR-carga-0.381Ni-carga-0.346V-carga-0.3V50-carga-0.325CarAro-carga+0.4 Penet-carga
1.09699	0.12189	0.62213	0.053S/C+0.637Trect+0.127D-carga-0.046CCR-carga+0.053Ni-carga-0.018V-carga-0.51V50-carga+0.515CarAro-carga+0.212Penet-carga

1.01873	0.11319	**0.73532**	-0.901S/C-0.3Trect+0.059D-carga-0.01CCR-carga+0.047Ni-carga-0.037V-carga-0.206V50-carga+0.214CarAro-carga+0.059Penet-carga
0.92162	0.1024	0.83773	-0.429S/C+0.701Trect-0.125D-carga-0.048CCR-carga-0.014Ni-carga+0.115V-carga+0.365V50-carga-0.389CarAro-carga-0.097Penet-carga
0.58351	0.06483	0.90256	0.039S/C-0.103Trect-0.205D-carga-0.401CCR-carga+0.387Ni-carga+0.753V-carga-0.202V50-carga-0.121CarAro-carga+0.111Penet-carga
0.41702	0.04634	0.9489	0.019S/C+0.039Trect-0.249D-carga-0.086CCR-carga+0.668Ni-carga-0.475V-carga-0.234V50-carga-0.137CarAro-carga-0.428Penet-carga
0.32112	0.03568	0.98458	-0.016S/C+0.022Trect-0.126D-carga+0.307CCR-carga-0.419Ni-carga+0.255V-carga-0.533V50-carga-0.21CarAro-carga-0.566Penet-carga

Eigenvectors

V1	V2	V3	V4	V5	V6	V7	
-0.0029	0.053	**-0.9006**	-0.4287	0.0391	0.0189	-0.0165	**S/C**
-0.0297	**0.6368**	-0.3001	0.7007	-0.1027	0.0388	0.0218	**Trect**
-0.4465	0.1269	0.0586	-0.1249	-0.2052	-0.2493	-0.1263	**D-carga**
-0.4222	-0.0461	-0.0099	-0.0478	-0.4012	-0.0858	0.3071	**CCR-carga**
-0.3815	0.053	0.0471	-0.0136	0.3873	0.6683	-0.419	**Ni-carga**
-0.3463	-0.0179	-0.0369	0.1146	0.7531	-0.4748	0.2553	**V-carga**
-0.3005	**-0.5099**	-0.2056	0.3652	-0.2017	-0.2343	-0.5333	**V50-carga**
-0.3254	**0.5152**	0.2143	-0.3889	-0.1215	-0.1374	-0.2098	**CarAro-carga**
0.3998	0.2121	0.0591	-0.0968	0.1112	-0.4276	-0.5655	**Penet-carga**

EL DETERMINANTE DE LA MATRIZ ES:

5.231906447837779E-4

COEFICIENTE DE ESFERICIDAD DE BARTLETT

1518.6684923879657

KMO

0.6006861424254947

MATRIZ ANTI-IMAGEN DE CORRELACIONES

S/C	Trect	D-carga	CCR-carga	Ni-carga	V-carga	V50-carga	CarAro-carga	Penet-carga
0.149(a)	-0.031	0.049	-0.052	0.027	-0.023	-0.044	-0.041	0.0060
-0.031	0.769(a)	0.012	0.0060	-0.0010	-0.054	-0.012	-0.034	-0.023
0.049	0.012	0.556(a)	-0.701	0.273	-0.534	-0.835	-0.941	0.215
-0.052	0.0060	-0.701	0.7(a)	-0.048	0.37	0.422	0.515	0.269
0.027	-0.0010	0.273	-0.048	0.756(a)	-0.362	-0.331	-0.401	0.451
-0.023	-0.054	-0.534	0.37	-0.362	0.649(a)	0.363	0.47	-0.027
-0.044	-0.012	-0.835	0.422	-0.331	0.363	0.415(a)	0.874	-0.149
-0.041	-0.034	-0.941	0.515	-0.401	0.47	0.874	0.408(a)	-0.296
0.0060	-0.023	0.215	0.269	0.451	-0.027	-0.149	-0.296	0.845(a)

3.5.3 Resultados técnica de *clustering* se utiliza la información de una serie de variables para cada sujeto u objeto y, conforme a estas variables se mide la similitud entre ellos. Una vez medida la similitud se agrupan en grupos homogéneos internamente y diferentes entre sí.

El algoritmo que usa el prototipo para esta técnica es el *simplekmeans*, los resultados que arroja son los siguientes: los centroides de los *cluster*, las instancias que se usaron en cada uno de ellos y la comparación de las medias así como cuáles variables hacen parte de un *cluster*. Estos elementos se pueden ver en la Figura 33 y a continuación su respectiva interpretación.

Figura 33. Resultados de *clustering*

```
Cluster analysis
----------------
S/C
C3
Trect
D-carga
CCR-carga
Ni-carga
V-carga
V50-carga
CarAro-carga
Penet-carga

kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 193.17358262247916

Cluster centroids:

Cluster 0
 Mean/Mode: 6.8181   9.6000   111.0909   1.0147   19.6194   122.2429   226.6885   45.5267   24.7514   48.6530
 Std Devs:  1.4154   0.0      7.6793     0.0088   1.9270    23.7980    42.1385    1.2731    2.3194    32.9165
```

Cluster 1

Mean/Mode:	6.5952	4.4216	112.6530	1.0128	19.0616	117.0795	240.1901	45.6611	24.2810	59.7938
Std Devs:	1.3815	2.9531	7.6594	0.0132	2.6768	23.3134	57.3586	2.5417	3.1338	59.8413

Clustered Instances

0 55 (27%)
1 147 (73%)

Comparación de medias:

	Cluster 0	Cluster 1
S/C	6.8181	6.5952
C3	9.6000	4.4216
Trect	111.0909	112.6530
D-carga	1.0147	1.0128
CCR-carga	19.6194	19.0616
Ni-carga	122.2429	117.0795
V-carga	226.6885	240.1901
V50-carga	45.5267	45.6611
CarAro-carga	24.7514	24.2810
Penet-carga	48.6530	59.7938

Se han agregado al *Dataset* nuevas variables con las segmentaciones de las variables originales para 2 clusters.

Sugerencias:

1. Con estas variables generadas a partir de la segmentación se aplica un Análisis de Regresión para estimar los comportamientos de las variables en los diferentes Cluster.

Fuente: Sistema de Predicción de Propiedades, SPP 2.0

El número de iteraciones para hallar la agrupación en este caso fue de 6 y la suma de los errores al cuadrado del algoritmo es de 193.17.

Number of iterations: 6
Within cluster sum of squared errors: 193.17358262247916

De un total de 205 registros, el *cluster0* tiene 55 y el *cluster1* 147. Es decir, se agruparon correctamente 202 registros.

Clustered Instances	
0	55 (27%)
1	147 (73%)

Para saber cuales variables están agrupadas en cada uno de los *cluster*, se tiene en cuenta el valor mayor, es decir, al *cluster0* pertenecen S/C, C3, D-carga, CCR-carga, CarAro-carga y al *cluster1* Trect, V-carga, V50-carga, y penet-carga.

Comparación de medias:

	Cluster 0	**Cluster 1**
S/C	**6.8181**	6.5952
C3	**9.6000**	4.4216
Trect	111.0909	**112.6530**
D-carga	**1.0147**	1.0128
CCR-carga	**19.6194**	19.0616
Ni-carga	**122.2429**	117.0795
V-carga	226.6885	**240.1901**
V50-carga	45.5267	**45.6611**
CarAro-carga	**24.7514**	24.2810
Penet-carga	48.6530	**59.7938**

3.5.4 Resultados de la técnica árboles de decisión esta técnica recibe como entrada una situación descrita por un conjunto de atributos y los clasifica respecto a una clase dependiendo del grado de entropía, que es el nivel de desorden en los valores del atributo y la ganancia, correspondiente a la cantidad de información que se gana al seleccionar un atributo, dando como resultado una estructura de clasificación en forma de árbol, donde las ramas están etiquetadas con los posibles valores de la prueba y las hojas representan los valores de la clase, el sumario del Algoritmo J48 y la Matriz de Confusión. Esto se puede observar en la Figura 34, y a continuación su respectiva explicación.

Figura 34. Resultados técnica Árboles de Decisión

```
J48 pruned tree
------------------

Penet-carga <= 38.9
|   Trect <= 100
|   |   V-carga <= 235.38
|   |   |   C3 <= 4.4217
|   |   |   |   S/C <= 5: Medio (2.0)
|   |   |   |   S/C > 5: Alto (4.0)
|   |   |   C3 > 4.4217: Bajo (6.0/1.0)
|   |   V-carga > 235.38: Bajo (9.0)
|   Trect > 100: Bajo (70.0/4.0)
Penet-carga > 38.9
|   Trect <= 100: Alto (33.0/1.0)
|   Trect > 100
|   |   S/C <= 6.5
|   |   |   S/C <= 5
|   |   |   |   Trect <= 115
|   |   |   |   |   C3 <= 4.4217
|   |   |   |   |   |   Ni-carga <= 122.01
|   |   |   |   |   |   |   Ni-carga <= 103.43: Alto (3.0/1.0)
|   |   |   |   |   |   |   Ni-carga > 103.43
|   |   |   |   |   |   |   |   CCR-carga <= 17.01: Bajo (2.0)
|   |   |   |   |   |   |   |   CCR-carga > 17.01: Medio (4.0/1.0)
|   |   |   |   |   |   Ni-carga > 122.01: Alto (2.0)
|   |   |   |   |   C3 > 4.4217: Bajo (3.0/1.0)
|   |   |   |   Trect > 115
|   |   |   |   |   Penet-carga <= 82.9: Bajo (5.0)
|   |   |   |   |   Penet-carga > 82.9: Medio (5.0/1.0)
|   |   |   S/C > 5
|   |   |   |   C3 <= 4.4217
|   |   |   |   |   Trect <= 115
|   |   |   |   |   |   Penet-carga <= 66.6
|   |   |   |   |   |   |   Ni-carga <= 119.6: Medio (2.0)
|   |   |   |   |   |   |   Ni-carga > 119.6: Alto (2.0)
|   |   |   |   |   |   Penet-carga > 66.6: Alto (10.0)
|   |   |   |   |   Trect > 115
|   |   |   |   |   |   Ni-carga <= 103.43: Alto (3.0)
|   |   |   |   |   |   Ni-carga > 103.43: Medio (6.0)
|   |   |   |   C3 > 4.4217
|   |   |   |   |   Ni-carga <= 131.25: Medio (4.0/1.0)
|   |   |   |   |   Ni-carga > 131.25: Bajo (2.0)
```

```
|   |   S/C > 6.5
|   |   |   C3 <= 4.4217: Alto (19.0)
|   |   |   C3 > 4.4217
|   |   |   |   Ni-carga <= 131.25: Alto (4.0)
|   |   |   |   Ni-carga > 131.25: Bajo (2.0)

Number of Leaves  :    23

Size of the tree : 45

=== Summary ===

Correctly Classified Instances         191               94.5545 %
Incorrectly Classified Instances        11                5.4455 %
Kappa statistic                          0.9088
Mean absolute error                      0.0608
Root mean squared error                  0.1743
Root relative squared error             38.872  %
Total Number of Instances              202

=== Confusion Matrix ===

  a  b  c   <-- classified as
 78  0  1 |  a = Alto
  1 93  2 |  b = Bajo
  1  6 20 |  c = Medio
```

Fuente: Sistema de Predicción de Propiedades, SPP 2.0

El árbol plano que se presenta puede interpretarse de la siguiente manera: el árbol describe la rama izquierda y después la derecha. Entonces si penet-carga es <=38.9 se va a trect si este es <= 100 se escoge V-carga con <=235.38 si se cumple C3 <= 4.4217, si este es positivo S/C <= 5 la variable dependiente es medio y si es >5 la variable dependiente es alto. Si C3 es >4.4217 el valor de la variable categórica es bajo. Y así sucesivamente con todo el árbol.

```
Penet-carga <= 38.9
|   Trect <= 100
|   |   V-carga <= 235.38
|   |   |   C3 <= 4.4217
|   |   |   |   S/C <= 5: Medio (2.0)
|   |   |   |   S/C > 5: Alto (4.0)
|   |   |   C3 > 4.4217: Bajo (6.0/1.0)
|   |   V-carga > 235.38: Bajo (9.0)
|   Trect > 100: Bajo (70.0/4.0)
Penet-carga > 38.9
|   Trect <= 100: Alto (33.0/1.0)
|   Trect > 100
|   |   S/C <= 6.5
|   |   |   S/C <= 5
|   |   |   |   Trect <= 115
|   |   |   |   |   C3 <= 4.4217
|   |   |   |   |   |   Ni-carga <= 122.01
|   |   |   |   |   |   |   Ni-carga <= 103.43: Alto (3.0/1.0)
|   |   |   |   |   |   |   Ni-carga > 103.43
```

```
|  |  |  |  |  |  |  |  |  CCR-carga <= 17.01: Bajo (2.0)
|  |  |  |  |  |  |  |  |  CCR-carga > 17.01: Medio (4.0/1.0)
|  |  |  |  |  |  |  |  Ni-carga > 122.01: Alto (2.0)
|  |  |  |  |  |  |  C3 > 4.4217: Bajo (3.0/1.0)
|  |  |  |  |  Trect > 115
|  |  |  |  |  |  Penet-carga <= 82.9: Bajo (5.0)
|  |  |  |  |  |  Penet-carga > 82.9: Medio (5.0/1.0)
|  |  |  |  S/C > 5
|  |  |  |  |  C3 <= 4.4217
|  |  |  |  |  |  Trect <= 115
|  |  |  |  |  |  |  Penet-carga <= 66.6
|  |  |  |  |  |  |  |  Ni-carga <= 119.6: Medio (2.0)
|  |  |  |  |  |  |  |  Ni-carga > 119.6: Alto (2.0)
|  |  |  |  |  |  |  Penet-carga > 66.6: Alto (10.0)
|  |  |  |  |  |  Trect > 115
|  |  |  |  |  |  |  Ni-carga <= 103.43: Alto (3.0)
|  |  |  |  |  |  |  Ni-carga > 103.43: Medio (6.0)
|  |  |  |  |  C3 > 4.4217
|  |  |  |  |  |  Ni-carga <= 131.25: Medio (4.0/1.0)
|  |  |  |  |  |  Ni-carga > 131.25: Bajo (2.0)
|  |  S/C > 6.5
|  |  |  C3 <= 4.4217: Alto (19.0)
|  |  |  C3 > 4.4217
|  |  |  |  Ni-carga <= 131.25: Alto (4.0)
|  |  |  |  Ni-carga > 131.25: Bajo (2.0)
```

El número de niveles de hojas (ramas sin hijos) es de 23 y el número de ramas del árbol es de 45.

```
Number of Leaves  :   23
Size of the tree : 45
```

En cuanto al sumario se muestra el número de instancias clasificadas correctamente, 191 y 11 incorrectamente clasificadas. El Estadístico de Kappa dice

Kappa	grado de acuerdo
< 0,00	sin acuerdo
>0,00 - 0,20	insignificante
0,21 - 0,40	discreto
>0,41 - 0,60	moderado
0,61 - 0,80	sustancial
0,81 - 1,00	casi perfecto

En el ejemplo se tienen un grado de aceptación casi perfecto ya que fue de 0.9088 y los errores son prácticamente muy bajos.

```
=== Summary ===
Correctly Classified Instances         191               94.5545 %
Incorrectly Classified Instances        11                5.4455 %
Kappa statistic                          0.9088
Mean absolute error                      0.0608
```

Root mean squared error	0.1743
Root relative squared error	38.872 %
Total Number of Instances	202

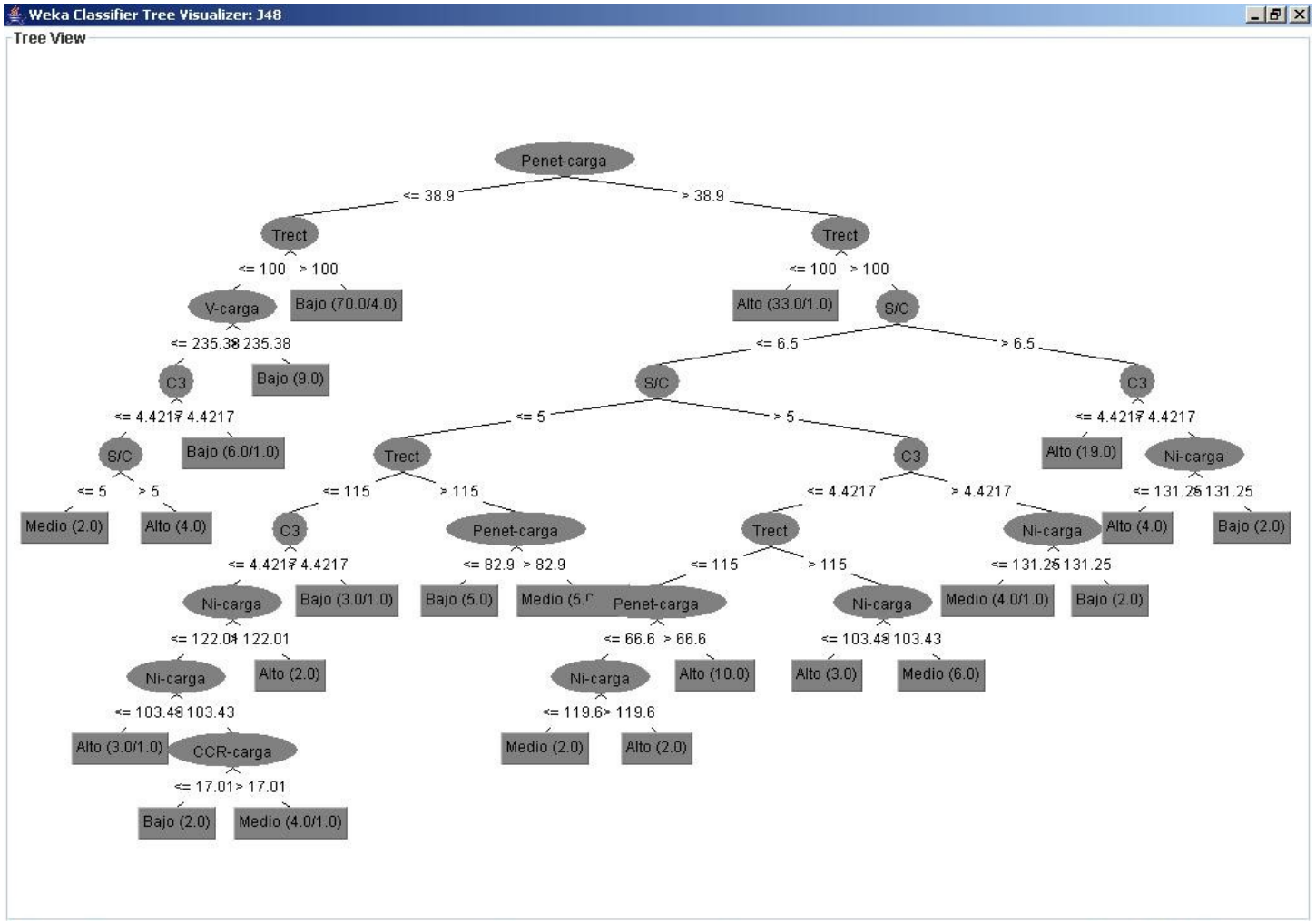
Y finalmente se tiene la Matriz de Confusión que es una matriz que contiene los números que reflejan la capacidad predictiva de la función discriminante. Se crea tabulando de forma cruzada el miembro del grupo concreto con el miembro del grupo predicho, los número de la diagonal representa clasificaciones correctas y los números fuera de la diagonal son clasificados incorrectamente.

Se concluye que en el primer renglón o para la categoría Alto se han clasificado 78 instancias en esta categoría y 1 en la categoría medio. Para la categoría bajo se ha clasificado 1 en alto, 93 en bajo y 2 en medio. Para la categoría medio se ha clasificado 1 en alto, 6 en bajo y 20 en medio. Esta clasificación no es muy buena ya que la mayor parte de los datos está clasificada en bajo.

```
=== Confusion Matrix ===
  a  b  c   <-- classified as
 78  0  1 |  a = Alto
  1 93  2 |  b = Bajo
  1  6 20 |  c = Medio
```

El árbol plano puede verse más claramente con la ayuda del árbol gráfico presentado en la Figura 35.

Figura 35. Árbol Gráfico



Fuente: Sistema de Predicción de Propiedades, SPP 2.0

3.5.5 Resultados de la técnica redes de bayes las Redes Bayesianas son una representación de un razonamiento probabilístico a partir de un conjunto de observaciones de diferentes variables, donde los nodos son las variables y los arcos representan la cardinalidad de las relaciones entre las variables. Esta técnica arroja en los resultados los siguientes elementos: probabilidad de las variables respecto a la dependiente, el sumario del algoritmo bayesnet y la matriz de confusión. Esto se puede observar en la Figura 36, y a continuación su respectiva explicación.

Figura 36. Resultados técnica redes de bayes

```
Bayes Network Classifier
----------------------
not using ADTree
#attributes=11 #classindex=10
Network structure (nodes followed by parents)

S/C(1): CatRendDMO
C3(1): CatRendDMO
Trect(2): CatRendDMO
D-carga(2): CatRendDMO
CCR-carga(2): CatRendDMO
Ni-carga(3): CatRendDMO
V-carga(2): CatRendDMO
V50-carga(3): CatRendDMO
CarAro-carga(2): CatRendDMO
Penet-carga(2): CatRendDMO
CatRendDMO(3):

Variables ordenadas por cardinalidad:

        Ni-carga        3
        V50-carga       3
        Trect           2
        D-carga         2
        CCR-carga       2
        V-carga         2
        CarAro-carga    2
        Penet-carga     2
        S/C             1
        C3              1

=== Summary ===

Correctly Classified Instances         152               75.2475 %
Incorrectly Classified Instances        50               24.7525 %
Kappa statistic                          0.5646
Mean absolute error                      0.1769
Root mean squared error                  0.3691
Root relative squared error             82.3106 %
Total Number of Instances              202

=== Confusion Matrix ===

  a  b  c   <-- classified as
```

```
71  8  0 |  a = Alto
15 81  0 |  b = Bajo
19  8  0 |  c = Medio
```

Fuente: Sistema de Predicción de Propiedades, SPP 2.0

La cardinalidad se puede leer de la siguiente manera

1. Variable al final del arco (después de :).
2. Valor de dependencia (Instancias en la que existe la relación).
3. Padre o variable antes del arco (antes de :).

Quiere decir que catRendDmo depende en 1 de S/C y así sucesivamente. Donde Ni-carga y V50-carga son los que más influyen.

```
S/C(1): CatRendDMO
C3(1): CatRendDMO
Trect(2): CatRendDMO
D-carga(2): CatRendDMO
CCR-carga(2): CatRendDMO
Ni-carga(3): CatRendDMO
V-carga(2): CatRendDMO
V50-carga(3): CatRendDMO
CarAro-carga(2): CatRendDMO
Penet-carga(2): CatRendDMO
CatRendDMO(3):

Variables ordenadas por cardinalidad:

        Ni-carga        3
        V50-carga       3
        Trect           2
        D-carga         2
        CCR-carga       2
        V-carga         2
        CarAro-carga    2
        Penet-carga     2
        S/C             1
        C3              1
```

En el sumario se tiene que fueron clasificadas 152 instancias correctamente y 50 incorrectamente. El Estadístico de Kappa con 0.5646 con un grado de acepción media y los errores bajos.

Kappa	**grado de acuerdo**
< 0,00	sin acuerdo
>0,00 - 0,20	insignificante
0,21 - 0,40	discreto
>0,41 - 0,60	moderado
0,61 - 0,80	sustancial
0,81 - 1,00	casi perfecto

```
=== Summary ===

Correctly Classified Instances         152               75.2475 %
Incorrectly Classified Instances        50               24.7525 %
Kappa statistic                          0.5646
Mean absolute error                      0.1769
Root mean squared error                  0.3691
Root relative squared error             82.3106 %
Total Number of Instances              202
```

Y finalmente se tiene la Matriz de Confusión que es una matriz que contiene los números que reflejan la capacidad predictiva de la función discriminante. Se crea tabulando de forma cruzada el miembro del grupo concreto con el miembro del grupo predicho, los número de la diagonal representa clasificaciones correctas y los números fuera de la diagonal son clasificados incorrectamente.

Se puede observar que 71 instancias fueron clasificadas en la categoría alta y 8 en bajo de la primera fila en la categoría alta. En la segunda fila en la categoría baja se clasificaron 15 en alto y 81 en bajo. En la tercera fila no se clasificaron instancias en la categoría medio, 19 en alto y 8 en bajo.

```
=== Confusion Matrix ===

a  b  c   <-- classified as
71  8  0 |  a = Alto
15 81  0 |  b = Bajo
19  8  0 |  c = Medio
```

3.5.6 Resultados GAMS la herramienta computacional GAMS (*General Algebraic Modeling System*) es un poderoso paquete matemático que permite entre muchas opciones, el modelamiento de sistemas lineales, no lineales y mixtos, de programación entera, y problemas de optimización. Este paquete ha sido diseñado para trabajar problemas de gran magnitud, y para ser usado desde computadoras personales, hasta *mainframes* y supercomputadoras.

Esta opción se usa cuando la técnica de regresión lineal no resuelve el modelo adecuadamente. Los resultados se pueden apreciar en la Figura 37.

Figura 37. Resultados Gams

```
Modelo: GRACE-1

ECUACIONES:
C3_Tr= b_C3*C3 + c_C3
CarAro_Tr= a_CarAro*CarAro-carga^2 + b_CarAro*CarAro-carga + c_CarAro
CCR_Tr= a_CCR*CCR-carga^2 + b_CCR*CCR-carga + c_CCR
iC4_Tr= b_iC4*i-C4 + c_iC4
```

```
Pen_Tr= a_Pen*Penet-carga^2 + b_Pen*Penet-carga + c_Pen
SC_Tr= a_SC*SC^2 + b_SC*SC + c_SC
Trect_Tr= a_Trect*Trect^2 + b_Trect*Trect + c_Trect
V50_Tr= a_V50*V50-carga^2 + b_V50*V50-carga - c_V50
SumTr = C3_Tr + CarAro_Tr + CCR_Tr + iC4_Tr + Pen_Tr + SC_Tr + Trect_Tr + V50_Tr
RendDMO = a_DMO*SumTr^2 + b_DMO*SumTr + c_DMO

VALOR DE LOS COEFICIENTES:
b_C3     =       1.07
c_C3     =       0.12
a_CarAro =       0.00
b_CarAro =      -0.09
c_CarAro =       1.16
a_CCR    =      -0.01
b_CCR    =       0.24
c_CCR    =       0.34
b_iC4    =      -0.35
c_iC4    =      13.33
a_Pen    =       0.00
b_Pen    =       0.00
c_Pen    =      -3.38
a_SC     =      -0.01
b_SC     =       0.12
c_SC     =      -2.17
a_Trect  =       0.00
b_Trect  =       0.06
c_Trect  =      11.79
a_V50    =       0.00
b_V50    =      -0.09
c_V50    =      21.19
a_DMO    =     -33.45
b_DMO    =      99.46
c_DMO    =      -0.02
MEDIDAS DE EFICIENCIA:
Grados de libertad  =      190
Suma de cuadrados   =    3861.13
Cuadrado medio      =      20.32
```

Fuente: Sistema de Predicción de Propiedades, SPP 2.0

El sistema de ecuaciones a resolver es de 9 x 9; como se puede apreciar son las siguientes,

```
C3_Tr= b_C3*C3 + c_C3
CarAro_Tr= a_CarAro*CarAro-carga^2 + b_CarAro*CarAro-carga + c_CarAro
CCR_Tr= a_CCR*CCR-carga^2 + b_CCR*CCR-carga + c_CCR
iC4_Tr= b_iC4*i-C4 + c_iC4
Pen_Tr= a_Pen*Penet-carga^2 + b_Pen*Penet-carga + c_Pen
SC_Tr= a_SC*SC^2 + b_SC*SC + c_SC
Trect_Tr= a_Trect*Trect^2 + b_Trect*Trect + c_Trect
V50_Tr= a_V50*V50-carga^2 + b_V50*V50-carga - c_V50
SumTr = C3_Tr + CarAro_Tr + CCR_Tr + iC4_Tr + Pen_Tr + SC_Tr + Trect_Tr + V50_Tr
RendDMO = a_DMO*SumTr^2 + b_DMO*SumTr + c_DMO
```

El software *Gams* resuelve este sistema y retorna los coeficientes que son

VALOR DE LOS COEFICIENTES:		
b_C3	=	1.07
c_C3	=	0.12
a_CarAro	=	0.00
b_CarAro	=	-0.09
c_CarAro	=	1.16
a_CCR	=	-0.01
b_CCR	=	0.24
c_CCR	=	0.34
b_iC4	=	-0.35
c_iC4	=	13.33
a_Pen	=	0.00
b_Pen	=	0.00
c_Pen	=	-3.38
a_SC	=	-0.01
b_SC	=	0.12
c_SC	=	-2.17
a_Trect	=	0.00
b_Trect	=	0.06
c_Trect	=	11.79
a_V50	=	0.00
b_V50	=	-0.09
c_V50	=	21.19
a_DMO	=	-33.45
b_DMO	=	99.46
c_DMO	=	-0.02

Y proporciona los grados de libertad la suma de cuadrados y el cuadrado medio que son:

MEDIDAS DE EFICIENCIA:		
Grados de libertad	=	190
Suma de cuadrados	=	3861.13
Cuadrado medio	=	20.32

3.5.7 Resultados ANOVA en la fase de análisis de varianza, se aplica ANOVA, una técnica estadística que sirve para determinar si las muestras provienen de poblaciones con igual media; ANOVA evita el aumento del error tipo 1 al comparar el conjunto de grupos en tratamiento, determinando si el conjunto completo de medidas muéstrales indica que las muestras fueron tomadas de la misma población general. En otras palabras, ANOVA es empleada para determinar la probabilidad de que las diferencias en las medidas entre varios grupos sean debidas únicamente al error muestral. A continuación se pueden observar los resultados obtenidos con el prototipo SPP 2.0:

Contrastes individuales de la t			
Variables	t Calculado	t Tabla	Resultado
Constante	3.26	1.96	*

S/C	7.80	1.96	*
C3	6.30	1.96	*
Trect	7.92	1.96	*
D-carga	1.26	1.96	-
CCR-carga	2.82	1.96	*
Ni-carga	1.04	1.96	-
V-carga	3.94	1.96	*
V50-carga	0.40	1.96	-
Penet-carga	2.31	1.96	*

*: Significativo
-: No significativo

Fuente: Sistema de Predicción de Propiedades, SPP 2.0

El estadístico *t* determina en el análisis de varianza, cuales variables son significativas para el modelo, de esta manera se calcula un estadístico *t* para cada variable, y un estadístico t para toda la Tabla, entonces los valores del estadístico *t* de las variables deben ser mayores a los valores del estadístico *t* de la Tabla, y bajo este criterio son significativos, de lo contrario son no significativos para el modelo.

3.6 EVALUACIÓN

En esta etapa se desarrolla la aplicación de los modelos propuestos iniciando con el modelo B (ver figura 30). Las variables son S/C, i-C4, Trect, D-carga, CCR-carga, Ni-carga, V-carga, V50-carga, CarAro-carga, Penet-carga, PIE-carga, N2b-carga, SatCarga, AromCarga, ResCarga, AsfCarga, FE-carga, S-carga, T10, T30, T50 y BCMI-carga.

Las variables independientes son 22, todas numéricas. Y el número de registros o la muestra de las variables es de 205. Para esta técnica no se tiene en cuenta la variable dependiente (sea nominal o categórica). Se procede con la fase de recogida y caracterización de datos. A continuación se aplica la técnica de Clusters (*SimpleKmeans*), con la cual se obtienen los siguientes resultados:

Cluster analysis

S/C
i-C4
Trect
D-carga
CCR-carga
Ni-carga
V-carga
V50-carga
CarAro-carga
Penet-carga

PIE-carga
N2b-carga
SatCarga
AromCarga
ResCarga
AsfCarga
FE-carga
S-carga
T10
T30
T50
BCMI-carga

kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 328.00272361738956

Cluster centroids:

Cluster 0
Mean/Mode: 6.8085 38.9700 111.3829 1.0158 19.6529 129.2961 233.2227 45.5844 25.0882 45.8706 891.3497 0.1944 14.5680 46.0319 28.7872 10.4659 0.2690 2.1167 507.6212 577.6063 650.7489 77.1111
Std Devs: 1.4201 0.0 7.7109 0.0078 1.9997 16.1855 33.1011 1.3137 2.0870 30.9301 24.0583 0.0094 2.3043 2.5144 2.9692 2.0886 0.0427 0.1142 13.4182 19.2466 34.3190 3.8016

Cluster 1
Mean/Mode: 6.6888 25.6919 112.0 1.0190 20.8442 110.4464 256.9688 46.6014 24.4064 35.6977 965.7555 0.1974 12.7666 43.4533 33.3311 10.3866 0.2965 2.1688 547.3311 622.4844 697.6044 76.0570
Std Devs: 1.3787 6.2439 7.5678 0.0155 2.4333 32.7904 88.2149 1.5869 3.1723 40.8706 17.2127 0.0148 2.0954 2.4012 3.4356 3.8430 0.0361 0.2677 6.9268 13.1331 20.4120 7.5013

Cluster 2
Mean/Mode: 6.5772 22.8209 112.6818 1.0099 18.3586 117.1549 229.5522 45.2420 24.1199 70.0299 872.1007 0.1908 16.3590 44.5827 28.8472 10.0100 0.2519 2.0666 498.1754 573.2245 649.9981 74.3770
Std Devs: 1.3899 1.8085 7.7423 0.0110 2.3569 19.8728 39.0307 2.6787 3.1172 62.3610 24.1781 0.0163 2.5837 2.7721 2.7445 1.9917 0.0360 0.1505 13.4260 15.8150 28.8005 5.5714

Clustered Instances

0 47 (23%)
1 45 (22%)
2 110 (54%)

Comparación de medias:

	Cluster 0	**Cluster 1**	**Cluster 2**
S/C	**6.8085**	6.6888	6.5772
i-C4	**38.9700**	25.6919	22.8209
Trect	111.3829	112.0	**112.6818**
D-carga	1.0158	**1.0190**	1.0099
CCR-carga	19.6529	**20.8442**	18.3586
Ni-carga	**129.2961**	110.4464	117.1549
V-carga	233.2227	**256.9688**	229.5522
V50-carga	45.5844	**46.6014**	45.2420
CarAro-carga	**25.0882**	24.4064	24.1199
Penet-carga	45.8706	35.6977	**70.0299**
PIE-carga	891.3497	**965.7555**	872.1007

N2b-carga	0.1944	**0.1974**	0.1908
SatCarga	14.5680	12.7666	**16.3590**
AromCarga	**46.0319**	43.4533	44.5827
ResCarga	28.7872	**33.3311**	28.8472
AsfCarga	**10.4659**	10.3866	10.0100
FE-carga	0.2690	**0.2965**	0.2519
S-carga	2.1167	**2.1688**	2.0666
T10	507.6212	**547.3311**	498.1754
T30	577.6063	**622.4844**	573.2245
T50	650.7489	**697.6044**	649.9981
BCMI-carga	77.1111	**76.0570**	74.3770

Se han agregado al *Dataset* nuevas variables con las segmentaciones de las variables originales para 3 clusters.

Sugerencias:

1. Con estas variables generadas a partir de la segmentación se aplica un Análisis de Regresión para estimar los comportamientos de las variables en los diferentes Cluster.

En los resultados se pueden apreciar la agrupación de las muestras por *cluster*, en este caso se han generado tres *clusters*, y en cada uno de ellos se pueden apreciar las medidas de los centroides por cada variable, y su respectiva desviación estándar para dar al usuario una idea de su agrupamiento; pero en lo que realmente hay que fijarse en estos resultados es en las instancias agrupadas o "*Clustered Instances*", las cuales muestran en cifras porcentuales las clasificaciones de la muestra en cada cluster, gracias a la aglomeración por centroides tomando en cuenta la distancia euclidiana entre ellos.

En la comparación de medias se puede ver la clasificación de variables por cluster, de tal forma de cómo quedan compuestos cada uno de ellos. Estas nuevas agrupaciones se encuentran en la sábana sobre la que se trabaja, en unas nuevas variables, cada *cluster* para el caso específico esta compuesto de:

Cluster0: S/C, i-C4, Ni-carga, CarAro-carga, AromCarga, AsfCarga.
Cluster1: D-carga, CCR-carga, V-carga, V50-carga, PIE-carga, N2b-carga, ResCarga, FE-carga, S-carga, T10, T30, T50, BCMI-carga.
Cluster2: Trect, Penet-carga, SatCarga.

Con los *clusters* ya conformados, se procede a hacer un análisis de regresión a cada uno de ellos, y así poder validar que estas nuevas variables hacen que el rendimiento sea mayor. Los resultados se pueden observar a continuación:

Para el cluster0:

Linear Regression Model

RendvDMO = 1.6441 * AromCarga-C0-R1 - 0.9085 * AsfCarga-C0-R1-+ 12.7792

Al evaluar el modelo remplazando las variables por los valores de la primera línea del *Dataset* se encontró que:

RendvDMO = 51.91270999999999

```
=== Summary ===
Correlation coefficient                  0.4827
Mean absolute error                      6.1421
Root mean squared error                  8.0974
Relative absolute error                 86.544  %
Root relative squared error             87.5805 %
Total Number of Instances               47
```

Para el cluster1:

```
Linear Regression Model
-----------------------
RendvDMO =-  0.0994 * V-carga-C1-R1 -  268.9411 * N2b-carga-C1-R1-+ 76.4926 * FE-carga-C1-R1 +
36.6535 * S-carga-C1-R1 + 75.1076

Al evaluar el modelo remplazando las variables por los valores de la primera línea del Dataset se encontró
que:

RendvDMO = 58.238803425773796

=== Summary ===
Correlation coefficient                  0.6622
Mean absolute error                      5.4502
Root mean squared error                  7.0629
Relative absolute error                 74.5662 %
Root relative squared error             74.9364 %
Total Number of Instances               45
```

Para el cluster 2:

```
Linear Regression Model
-----------------------
RendvDMO = 0.0423 * Penet-carga-C2-R1 + 49.1331

Al evaluar el modelo remplazando las variables por los valores de la primera línea del Dataset se encontró
que:

RendvDMO = 62.6267

=== Summary ===
Correlation coefficient                  0.3003
Mean absolute error                      6.5679
Root mean squared error                  8.3345
Relative absolute error                 98.0331 %
Root relative squared error             95.3852 %
Total Number of Instances               110
```

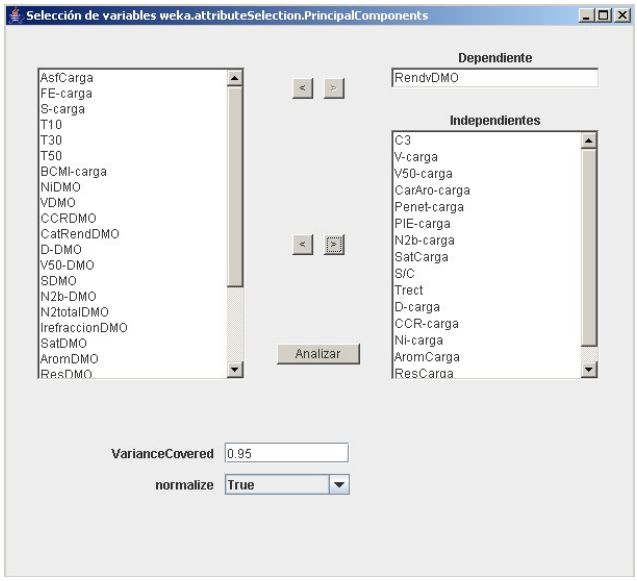
Del análisis anterior se puede concluir que sólo el agrupamiento del *cluster1* es aceptable para otros análisis con las variables que este contiene, ya que da un rendimiento relativamente alto con una correlación aceptable, en el *cluster0* el rendimiento es similar al del *cluster1* pero el coeficiente de correlación entre las variables es muy bajo, entonces esta agrupación es inaceptable definitivamente, y finalmente el *cluster2* ofrece un alto rendimiento para la variable dependiente, pero

igual que en el *cluster0,* el coeficiente de correlación entre las variables es muy bajo, lo cual lo hace una agrupación inaceptable también.

Para el modelo A (ver figura 29). Las variables son S/C, C3, i-C4, n-C4, Trect, D-carga, CCR-carga, Ni-carga, V-carga, V50-carga, CarAro-carga, Penet-carga, PIE-carga, N2b-carga, SatCarga, AromCarga, ResCarga, AsfCarga, FE-carga, S-carga, T10, T30, T50, BCMI-carga. La variable dependiente se clasifica en catRendDMO (medio, bajo y alto) y numérica con RendvDMO. Las variables independientes son 24, todas numéricas, y el número de registros o la muestra de las variables es de 205.

Siguiendo el orden jerárquico presentado en la fase de modelado, se aplica el *modelo A*. iniciando con la técnica de Análisis de Componentes Principales. A través de varios ejemplos se puede concluir que C3, i-C4, n-C4 debido a los valores que toman (dos valores y son constantes), no pueden ir juntas en un análisis ya que los resultados no se pueden obtener. También que las variables carAro-carga y Arom-carga no son tan importantes, es decir, no influyen mucho en la predicción del rendimiento. Esto se puede ver en el Anexo D. Se usaron las variables S/C, C3, Trect, D-carga, CCR-carga, Ni-carga, V-carga, V50-carga, CarAro-carga, Penet-carga, PIE-carga, N2b-carga, SatCarga, AromCarga, ResCarga y en las opciones varianza explicada de 0.95 y normalizados los resultados, como se ve en la Figura 38. A continuación se presentan algunos ejemplos y se define un conjunto de datos que cumple con todas las condiciones del análisis de componentes principales.

Figura 38. Selección de variables



Fuente: Sistema de Predicción de Propiedades, SPP 2.0

Principal Components Attribute Transformer

Correlation matrix

1	0.07	0.03	0	0.02	-0.01	0	0.02	-0.01	-0.01	0.02	-0.02	0	-0.01	0
0.07	1	-0.09	0.07	0.1	0.1	-0.11	-0.03	0.07	-0.09	0.07	0.02	-0.14	0.18	0.01
0.03	-0.09	1	0.06	0.02	0.05	0.07	-0.01	0.09	0	-0.04	0.03	0.01	0	-0.05
0	0.07	0.06	1	0.88	0.68	0.63	0.55	0.82	-0.71	0.21	0.38	-0.72	0.27	-0.18
0.02	0.1	0.02	**0.88**	1	0.59	0.52	0.57	0.59	-0.79	0.34	0.37	-0.75	0.11	-0.06
-0.01	0.1	0.05	**0.68**	**0.59**	1	0.6	0.43	0.56	-0.67	-0.15	0.37	-0.52	0.43	-0.4
0	-0.11	0.07	**0.63**	**0.52**	**0.6**	1	0.44	0.4	-0.55	0.15	0.3	-0.33	-0.02	-0.17
0.02	-0.03	-0.01	**0.55**	**0.57**	**0.43**	**0.44**	1	0.05	-0.59	0.21	0.26	-0.52	0.12	0.02
-0.01	0.07	0.09	**0.82**	**0.59**	**0.56**	**0.4**	0.05	1	-0.38	-0.11	0.2	-0.41	0.31	-0.37
-0.01	-0.09	0	**-0.71**	**-0.79**	**-0.67**	**-0.55**	**-0.59**	**-0.38**	1	-0.24	-0.55	0.76	-0.12	-0.08
0.02	0.07	-0.04	0.21	**0.34**	-0.15	0.15	0.21	-0.11	-0.24	1	0.25	-0.49	-0.16	0.54
-0.02	0.02	0.03	**0.38**	**0.37**	**0.37**	**0.3**	0.26	0.2	**-0.55**	0.25	1	-0.57	0.03	0.23
0	-0.14	0.01	**-0.72**	**-0.75**	**-0.52**	**-0.33**	**-0.52**	**-0.41**	**0.76**	**-0.49**	**-0.57**	1	-0.29	-0.28
-0.01	0.18	0	0.27	0.11	**0.43**	-0.02	0.12	0.31	-0.12	-0.16	0.03	-0.29	1	-0.6
0	0.01	-0.05	-0.18	-0.06	**-0.4**	-0.17	0.02	**-0.37**	-0.08	**0.54**	0.23	-0.28	**-0.6**	1

Sugerencia:

Se recomienda trabajar con las variables de los componentes cuyos eigenvalues sean superiores a la unidad.

eigenvalue	proportion	cumulative	
5.50968	0.36731	0.36731	-0.002S/C-0.043C3-0.02Trect-0.398D-carga-0.381CCR-carga-0.336Ni-carga-0.286V-carga-0.27V50-carga-0.278CarAro-carga+0.371Penet-carga-0.113PIE-carga-0.234N2b-carga+0.354SatCarga-0.128AromCarga+0.046ResCarga
2.40437	0.16029	0.5276	-0.012S/C+0.015C3+0.062Trect+0.071D-carga-0.061CCR-carga+0.249Ni-carga+0.035V-carga-0.117V50-carga+0.279CarAro-carga+0.12 Penet-carga-0.475PIE-carga-0.212N2b-carga+0.213SatCarga+0.387AromCarga-0.593ResCarga
1.2792	0.08528	0.61288	-0.112S/C-0.683C3+0.339Trect+0.041D-carga+0.004CCR-carga+0.022Ni-carga+0.394V-carga+0.097V50-carga+0.014CarAro-carga-0.011Penet-carga-0.139PIE-carga-0.014N2b-carga+0.223SatCarga-0.411AromCarga-0.029ResCarga
1.03854	0.06924	0.68212	-0.794S/C-0.161C3-0.536Trect-0.051D-carga-0.044CCR-carga+0.06 Ni-carga+0.001V-carga+0.133V50-carga-0.127CarAro-carga-0.042Penet-carga-0.061PIE-carga+0.037N2b-carga-0.016SatCarga+0.082AromCarga-0.031ResCarga
0.97926	0.06528	0.7474	-0.515S/C+0.071C3+0.577Trect+0.047D-carga-0.026CCR-carga-0.066Ni-carga-0.209V-carga-0.402V50-carga+0.277CarAro-carga+0.075Penet-carga+0.084PIE-carga+0.226N2b-carga-0.144SatCarga+0.087AromCarga+0.123ResCarga
0.88861	0.05924	0.80664	-0.021S/C+0.144C3-0.391Trect+0.233D-carga+0.187CCR-carga-0.134Ni-carga+0.15 V-carga-0.421V50-carga+0.501CarAro-carga+0.099Penet-carga+0.063PIE-carga-0.226N2b-carga+0.129SatCarga-0.429AromCarga+0.06 ResCarga
0.79201	0.0528	0.85945	0.177S/C+0.006C3-0.214Trect-0.156D-carga-0.216CCR-carga+0.228Ni-carga+0.034V-carga-0.308V50-carga+0.026CarAro-carga-0.152Penet-carga-0.398PIE-carga+0.697N2b-carga+0.035SatCarga-0.154AromCarga+0.1 ResCarga
0.7391	0.04927	0.90872	-0.242S/C+0.667C3+0.248Trect-0.079D-carga+0.006CCR-carga+0.157Ni-carga+0.246V-carga+0.198V50-carga-0.198CarAro-carga-0.144Penet-carga-0.23PIE-carga-0.109N2b-carga+0.229SatCarga-0.356AromCarga+0.029ResCarga
0.52507	0.035	0.94372	-0.015S/C+0.108C3-0.022Trect-0.109D-carga-0.214CCR-carga+0.076Ni-carga+0.616V-carga-0.254V50-carga-0.149CarAro-carga+0.171Penet-carga+0.543PIE-carga+0.163N2b-carga+0.124SatCarga+0.212AromCarga-0.217ResCarga
0.32404	0.0216	0.96533	-0.018S/C+0.133C3-0.016Trect+0.221D-carga+0.044CCR-carga-0.454Ni-carga-0.057V-carga+0.431V50-carga+0.148CarAro-carga+0.38 Penet-carga+0.009PIE-carga+0.497N2b-carga+0.22 SatCarga-0.072AromCarga-0.262ResCarga

Eigenvectors

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	
-0.0018	-0.0123	-0.1118	-0.7941	-0.5151	-0.0209	0.1769	-0.2424	-0.0155	-0.0179	S/C
-0.043	0.0145	-0.6834	-0.1606	0.071	0.1442	0.0058	0.6671	0.1082	0.133	C3
-0.0199	0.0615	0.3387	-0.5357	0.5771	-0.3907	-0.2143	0.2477	-0.0222	-0.0157	Trect
-0.398	0.0714	0.0406	-0.051	0.0468	0.233	-0.1561	-0.0791	-0.1087	0.221	D-carga
-0.3814	-0.0609	0.0045	-0.0438	-0.0264	0.1873	-0.2164	0.0062	-0.2144	0.0435	CCR-carga
-0.336	0.2491	0.0219	0.0596	-0.066	-0.1344	0.2282	0.1572	0.0764	-0.4545	Ni-carga
-0.2859	0.0348	0.3939	0.001	-0.2086	0.1499	0.0338	0.2457	0.6158	-0.0573	V-carga
-0.2705	-0.1174	0.0972	0.1332	-0.4021	-0.4206	-0.3085	0.1981	-0.2542	0.4307	V50-carga
-0.2778	0.2794	0.0135	-0.1268	0.2771	0.5006	0.0264	-0.1976	-0.1492	0.1477	CarAro-carga
0.3708	0.1196	-0.0107	-0.0424	0.0748	0.0992	-0.152	-0.144	0.1705	0.38	Penet-carga
-0.113	-0.475	-0.1386	-0.0607	0.0841	0.0635	-0.3982	-0.2296	0.5427	0.0095	PIE-carga
-0.2343	-0.2119	-0.0142	0.0368	0.2263	-0.2257	0.6971	-0.1086	0.1627	0.4972	N2b-carga
0.3541	0.2131	0.223	-0.0156	-0.1442	0.1285	0.0347	0.2289	0.1241	0.22	SatCarga
-0.1277	0.3869	-0.4106	0.0817	0.0875	-0.4287	-0.1539	-0.3559	0.2122	-0.0717	AromCarga
0.0459	-0.5933	-0.0292	-0.0312	0.1229	0.06	0.1002	0.0287	-0.2174	-0.2621	ResCarga

EL DETERMINANTE DE LA MATRIZ ES:

3.072345685232958E-7

COEFICIENTE DE ESFERICIDAD DE BARTLETT

2984.1352085374538

KMO

0.4883132341952253

MATRIZ ANTI-IMAGEN DE CORRELACIONES

S/C	C3	Trect	D-carga	CCR-carga	Ni-carga	V-carga	V50-carga	CarAro-carga	Penet-carga	PIE-carga	N2b-carga	SatCarga	AromCarga	ResCarga
0.181(a)	-0.08	-0.04	0.076	-0.052	0.03	-0.056	-0.077	-0.073	0.011	-0.047	0.0030	-0.0060	-0.019	-0.013
-0.08	**0.399(a)**	0.085	-0.036	-0.106	-0.174	0.096	0.071	0.035	0.0070	-0.01	-0.0070	-0.193	-0.194	-0.185
-0.04	0.085	**0.701(a)**	-0.0030	-0.0080	-0.024	-0.015	0.0020	-0.0080	-0.032	0.0070	-0.04	-0.029	-0.015	-0.021
0.076	-0.036	-0.0030	**0.515(a)**	-0.737	0.109	-0.752	-0.935	-0.978	-0.036	-0.48	-0.221	-0.204	-0.482	-0.431
-0.052	-0.106	-0.0080	-0.737	**0.485(a)**	0.024	0.68	0.616	0.672	0.319	0.105	0.385	0.609	0.806	0.778
0.03	-0.174	-0.024	0.109	0.024	**0.873(a)**	-0.32	-0.102	-0.11	0.268	0.221	-0.034	0.285	0.04	0.215
-0.056	0.096	-0.015	-0.752	0.68	-0.32	**0.419(a)**	0.631	0.695	0.191	0.083	0.182	0.194	0.562	0.487
-0.077	0.071	0.0020	-0.935	0.616	-0.102	0.631	**0.348(a)**	0.954	0.026	0.524	0.228	0.212	0.414	0.384
-0.073	0.035	-0.0080	-0.978	0.672	-0.11	0.695	0.954	**0.367(a)**	-0.024	0.539	0.213	0.236	0.467	0.433
0.011	0.0070	-0.032	-0.036	0.319	0.268	0.191	0.026	-0.024	**0.893(a)**	-0.238	0.262	0.0020	0.129	0.203
-0.047	-0.01	0.0070	-0.48	0.105	0.221	0.083	0.524	0.539	-0.238	**0.501(a)**	0.033	0.113	0.03	-0.024

0.0030	-0.0070	-0.04	-0.221	0.385	-0.034	0.182	0.228	0.213	0.262	0.033	**0.634(a)**	0.4	0.358	0.305
-0.0060	-0.193	-0.029	-0.204	0.609	0.285	0.194	0.212	0.236	0.0020	0.113	0.4	**0.588(a)**	0.827	0.851
-0.019	-0.194	-0.015	-0.482	0.806	0.04	0.562	0.414	0.467	0.129	0.03	0.358	0.827	**0.212(a)**	0.929
-0.013	-0.185	-0.021	-0.431	0.778	0.215	0.487	0.384	0.433	0.203	-0.024	0.305	0.851	0.929	**0.266(a)**

Teniendo en cuenta los criterios de esta técnica se observa que solo 40 datos por debajo de la diagonal de 84 son mayores de 0.3, indica que no son significativos, el KMO es menor de 0.4 lo que significa es que no es aceptable hacer el análisis y 3 variables de la diagonal principal son menores de 0.3. Eliminado la menor de ellas sin contar S/C ya que es una variable importante para el investigador, se eliminará entonces Arom-carga y se tiene que el KMO sube a 0.6 y es aceptable el análisis de componentes principales, las otras variables son mayores de 0.3 en la diagonal principal. Como se aprecia continuación.

Principal Components Attribute Transformer

Correlation matrix

1	0.07	0.03	0	0.02	-0.01	0	0.02	-0.01	-0.01	0.02	-0.02	0	0	
0.07	1	-0.09	0.07	0.1	0.1	-0.11	-0.03	0.07	-0.09	0.07	0.02	-0.14	0.01	
0.03	-0.09	1	0.06	0.02	0.05	0.07	-0.01	0.09	0	-0.04	0.03	0.01	-0.05	
0	0.07	0.06	1	0.88	0.68	0.63	0.55	0.82	-0.71	0.21	0.38	-0.72	-0.18	
0.02	0.1	0.02	**0.88**	1	0.59	0.52	0.57	0.59	-0.79	0.34	0.37	-0.75	-0.06	
-0.01	0.1	0.05	**0.68**	**0.59**	1	0.6	0.43	0.56	-0.67	-0.15	0.37	-0.52	-0.4	
0	-0.11	0.07	**0.63**	**0.52**	**0.6**	1	0.44	0.4	-0.55	0.15	0.3	-0.33	-0.17	
0.02	-0.03	-0.01	**0.55**	**0.57**	**0.43**	**0.44**	1	0.05	-0.59	0.21	0.26	-0.52	0.02	
-0.01	0.07	0.09	**0.82**	**0.59**	**0.56**	**0.4**	0.05	1	-0.38	-0.11	0.2	-0.41	-0.37	
-0.01	-0.09	0	**-0.71**	**-0.79**	**-0.67**	**-0.55**	**-0.59**	**-0.38**	1	-0.24	-0.55	0.76	-0.08	
0.02	0.07	-0.04	0.21	**0.34**	-0.15	0.15	0.21	-0.11	-0.24	1	0.25	-0.49	0.54	
-0.02	0.02	0.03	**0.38**	**0.37**	**0.37**	**0.3**	0.26	0.2	**-0.55**	0.25	1	-0.57	0.23	
0	-0.14	0.01	**-0.72**	**-0.75**	**-0.52**	**-0.33**	**-0.52**	**-0.41**	**0.76**	**-0.49**	**-0.57**	1	-0.28	
0	0.01	-0.05	-0.18	-0.06	**-0.4**	-0.17	0.02	**-0.37**	-0.08	**0.54**	0.23	-0.28	1	

Sugerencia:

Se recomienda trabajar con las variables de los componentes cuyos eigenvalues sean superiores a la unidad.

eigenvalue	proportion	cumulative	
5.43732	0.38838	0.38838	-0.002S/C-0.039C3-0.02Trect-0.4D-carga-0.387CCR-carga-0.33Ni-carga-0.293V-carga-0.275V50-carga-0.274CarAro-carga+0.377Penet-carga-0.124PIE-carga-0.241N2b-carga+0.357SatCarga+0.027ResCarga
2.15859	0.15419	0.54257	-0.013S/C-0.048C3+0.094Trect+0.114D-carga-0.021CCR-carga+0.278Ni-carga+0.129V-carga-0.11V50-carga+0.332CarAro-carga+0.094Penet-carga-0.522PIE-carga-0.217N2b-carga+0.257SatCarga-0.603ResCarga
1.1553	0.08252	0.62509	-0.256S/C-0.789C3+0.382Trect-0.045D-carga-0.066CCR-carga-0.026Ni-carga+0.274V-carga+0.194V50-carga-0.168CarAro-carga-0.012Penet-carga+0.002PIE-carga+0.079N2b-carga+0.094SatCarga+0.048ResCarga
1.03411	0.07386	0.69895	0.735S/C+0.038C3+0.653Trect+0.028D-carga+0.017CCR-carga-0.068Ni-carga-0.002V-carga-0.098V50-carga+0.076CarAro-carga+0.044Penet-carga+0.084PIE-carga+0.004N2b-carga-0.004SatCarga+0.046ResCarga

0.97607 0.06972 0.76867 0.548S/C-0.155C3-0.463Trect-0.076D-carga0 CCR-carga+0.065Ni-carga+0.2 V-carga+0.46 V50-carga-0.347CarAro-carga-0.078Penet-carga-0.065PIE-carga-0.197N2b-carga+0.123SatCarga-0.12ResCarga

0.81696 0.05835 0.82702 -0.037S/C+0.279C3+0.202Trect-0.269D-carga-0.239CCR-carga+0.311Ni-carga-0.071V-carga+0.15 V50-carga-0.356CarAro-carga-0.217Penet-carga-0.412PIE-carga+0.531N2b-carga-0.003SatCarga+0.004ResCarga

0.77792 0.05557 0.88259 -0.3S/C+0.401C3+0.402Trect+0.01 D-carga+0.121CCR-carga-0.026Ni-carga+0.009V-carga+0.461V50-carga-0.257CarAro-carga-0.001Penet-carga+0.118PIE-carga-0.518N2b-carga+0.048SatCarga-0.096ResCarga

0.56381 0.04027 0.92286 0.047S/C-0.281C3+0.021Trect+0.087D-carga+0.168CCR-carga-0.073Ni-carga-0.77V-carga+0.276V50-carga+0.089CarAro-carga-0.071Penet-carga-0.328PIE-carga-0.102N2b-carga-0.276SatCarga+0.027ResCarga

0.35425 0.0253 0.94817 0.007S/C+0.016C3+0 Trect-0.016D-carga-0.04CCR-carga+0.092Ni-carga+0.22 V-carga-0.162V50-carga+0.104CarAro-carga-0.318Penet-carga-0.439PIE-carga-0.419N2b-carga+0.012SatCarga+0.659ResCarga

0.3096 0.02211 0.97028 -0.011S/C+0.168C3-0.022Trect+0.277D-carga-0.091CCR-carga-0.484Ni-carga+0.166V-carga+0.435V50-carga+0.28 CarAro-carga+0.308Penet-carga-0.287PIE-carga+0.293N2b-carga+0.221SatCarga+0.212ResCarga

Eigenvectors

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	
-0.0023	-0.0134	-0.2561	0.7349	0.5479	-0.0367	-0.2998	0.0468	0.0072	-0.0113	S/C
-0.0389	-0.0475	-0.7892	0.0384	-0.1551	0.2787	0.4013	-0.2814	0.0164	0.1681	C3
-0.0199	0.0938	0.3823	0.653	-0.4635	0.2024	0.4016	0.0211	0.0002	-0.0223	Trect
-0.3995	0.1145	-0.0455	0.0278	-0.076	-0.2694	0.0103	0.0873	-0.0159	0.2766	D-carga
-0.3873	-0.0209	-0.0665	0.0166	-0.0003	-0.239	0.1211	0.1684	-0.0404	-0.0914	CCR-carga
-0.3304	0.2784	-0.0256	-0.0675	0.0652	0.3112	-0.0263	-0.0727	0.0925	-0.4841	Ni-carga
-0.2926	0.1291	0.2737	-0.0017	0.1996	-0.071	0.0089	-0.7699	0.2201	0.166	V-carga
-0.2746	-0.1095	0.1938	-0.098	0.4601	0.1495	0.4606	0.2757	-0.1623	0.4347	V50-carga
-0.2736	0.3317	-0.1685	0.076	-0.3469	-0.3562	-0.2573	0.0889	0.1039	0.2799	CarAro-carga
0.3771	0.0943	-0.0116	0.0435	-0.0782	-0.2174	-0.001	-0.0713	-0.3176	0.308	Penet-carga
-0.124	-0.5222	0.0021	0.0839	-0.0649	-0.4123	0.1185	-0.3284	-0.4389	-0.2873	PIE-carga
-0.2407	-0.2168	0.079	0.0038	-0.1975	0.5309	-0.518	-0.1015	-0.4192	0.2931	N2b-carga
0.3568	0.2575	0.094	-0.0044	0.1234	-0.0034	0.0476	-0.2762	0.0115	0.2208	SatCarga
0.027	-0.6034	0.048	0.0458	-0.1203	0.0036	-0.0963	0.0273	0.659	0.2121	ResCarga

EL DETERMINANTE DE LA MATRIZ ES:

4.4324593221332814E-6

COEFICIENTE DE ESFERICIDAD DE BARTLETT

2452.9846392685463

KMO

0.6024608974224358

MATRIZ ANTI-IMAGEN DE CORRELACIONES

S/C	C3	Trect	D-carga	CCR-carga	Ni-carga	V-carga	V50-carga	CarAro-carga	Penet-carga	PIE-carga	N2b-carga	SatCarga	ResCarga
0.172(a)	-0.085	-0.04	0.076	-0.062	0.03	-0.054	-0.076	-0.072	0.013	-0.047	0.011	0.017	0.012

-0.085	**0.31(a)**	0.084	-0.151	0.086	-0.17	0.253	0.17	0.145	0.033	-0.0040	0.067	-0.059	-0.013
-0.04	0.084	**0.715(a)**	-0.012	0.0070	-0.023	-0.0080	0.01	-0.0010	-0.03	0.0080	-0.037	-0.029	-0.019
0.076	-0.151	-0.012	**0.547(a)**	-0.673	0.147	-0.664	-0.922	-0.972	0.029	-0.531	-0.059	0.394	0.052
-0.062	0.086	0.0070	-0.673	**0.7(a)**	-0.014	0.464	0.523	0.565	0.365	0.135	0.174	-0.174	0.132
0.03	-0.17	-0.023	0.147	-0.014	**0.768(a)**	-0.416	-0.131	-0.146	0.265	0.22	-0.053	0.447	0.482
-0.054	0.253	-0.0080	-0.664	0.464	-0.416	**0.505(a)**	0.529	0.591	0.144	0.08	-0.024	-0.582	-0.113
-0.076	0.17	0.01	-0.922	0.523	-0.131	0.529	**0.388(a)**	0.945	-0.03	0.562	0.094	-0.255	-0.0020
-0.072	0.145	-0.0010	-0.972	0.565	-0.146	0.591	0.945	**0.4(a)**	-0.097	0.593	0.056	-0.302	-0.0020
0.013	0.033	-0.03	0.029	0.365	0.265	0.144	-0.03	-0.097	**0.883(a)**	-0.244	0.233	-0.188	0.224
-0.047	-0.0040	0.0080	-0.531	0.135	0.22	0.08	0.562	0.593	-0.244	**0.447(a)**	0.024	0.157	-0.145
0.011	0.067	-0.037	-0.059	0.174	-0.053	-0.024	0.094	0.056	0.233	0.024	**0.896(a)**	0.198	-0.078
0.017	-0.059	-0.029	0.394	-0.174	0.447	-0.582	-0.255	-0.302	-0.188	0.157	0.198	**0.732(a)**	0.398
0.012	-0.013	-0.019	0.052	0.132	0.482	-0.113	-0.0020	-0.0020	0.224	-0.145	-0.078	0.398	**0.612(a)**

Otro tipo de ejemplos como estos pueden verse en el Anexo D. Como la variable S/C es muy importante y no puede ser eliminada, se decide escoger otras variables y realizar de nuevo este análisis. Estas son S/C, C3, Trect, D-carga, CCR-carga, Ni-carga, V-carga, V50-carga, Penet-carga, PIE-carga, N2b-carga, SatCarga, ResCarga, BCMI-carga, con las opciones de varianza aceptada de un 0.95 y los datos normalizados.

Y con estos datos se obtuvo un buen resultado, es decir, existen 40 de 84 que son significativos en la matriz de correlación, el KMO es de 0.76, el determinante de la matriz es bajo y la diagonal principal en la matriz anti-imagen de correlaciones es toda mayor de 0.3, se obtienen cuatro eigenvalores con una varianza explicada de 71%. Como se aprecia a continuación:

Principal Components Attribute Transformer

Correlation matrix

1	0.07	0.03	0	0.01	-0.01	0	0.01	-0.01	0.02	-0.02	0	0	0
0.07	1	-0.09	0.06	0.08	0.09	-0.12	-0.03	-0.08	0.07	0.02	-0.13	0.02	0.07
0.03	-0.09	1	0.06	0.03	0.05	0.07	-0.01	-0.01	-0.04	0.03	0	-0.06	0.06
0	0.06	0.06	1	0.87	0.68	0.63	0.55	-0.72	0.2	0.37	-0.72	-0.2	0.95
0.01	0.08	0.03	**0.87**	1	0.59	0.54	0.57	-0.79	0.32	0.35	-0.73	-0.09	0.78
-0.01	0.09	0.05	**0.68**	**0.59**	1	0.6	0.43	-0.67	-0.15	0.37	-0.52	-0.41	0.73
0	-0.12	0.07	**0.63**	**0.54**	**0.6**	1	0.44	-0.56	0.14	0.29	-0.33	-0.19	0.54
0.01	-0.03	-0.01	**0.55**	**0.57**	**0.43**	**0.44**	1	-0.59	0.21	0.26	-0.52	0.01	0.48
-0.01	-0.08	-0.01	**-0.72**	**-0.79**	**-0.67**	**-0.56**	**-0.59**	1	-0.24	-0.54	0.75	-0.07	-0.65
0.02	0.07	-0.04	0.2	**0.32**	-0.15	0.14	0.21	-0.24	1	0.25	-0.49	0.54	-0.01
-0.02	0.02	0.03	**0.37**	**0.35**	**0.37**	**0.29**	0.26	**-0.54**	0.25	1	-0.57	0.23	0.35
0	-0.13	0	**-0.72**	**-0.73**	**-0.52**	**-0.33**	**-0.52**	**0.75**	**-0.49**	**-0.57**	1	-0.28	-0.64
0	0.02	-0.06	-0.2	-0.09	**-0.41**	-0.19	0.01	-0.07	**0.54**	0.23	-0.28	1	-0.37
0	0.07	0.06	**0.95**	**0.78**	**0.73**	**0.54**	**0.48**	**-0.65**	-0.01	**0.35**	**-0.64**	**-0.37**	1

Sugerencia:
Se recomienda trabajar con las variables de los componentes cuyos eigenvalues sean superiores a la unidad.

eigenvalue	proportion	cumulative	
5.79497	0.41393	0.41393	0.002S/C+0.031C3+0.02 Trect+0.385D-carga+0.372CCR-carga+0.323Ni-carga+0.284V-carga+0.279V50-carga-0.366Penet-carga+0.112PIE-carga+0.228N2b-carga-0.343SatCarga-0.04ResCarga+0.362BCMI-carga
2.08685	0.14906	0.56299	0.014S/C+0.075C3-0.092Trect-0.087D-carga+0.023CCR-carga-0.283Ni-carga-0.137V-carga+0.054V50-carga-0.089Penet-carga+0.545PIE-carga+0.237N2b-carga-0.28SatCarga+0.624ResCarga-0.218BCMI-carga
1.14432	0.08174	0.64472	0.271S/C+0.783C3-0.453Trect+0.022D-carga+0.043CCR-carga+0.072Ni-carga-0.243V-carga-0.096V50-carga-0.01Penet-carga-0.058PIE-carga-0.093N2b-carga-0.07SatCarga-0.097ResCarga+0.079BCMI-carga
1.03122	0.07366	0.71838	-0.812S/C-0.061C3-0.574Trect-0.014D-carga-0.018CCR-carga+0.039Ni-carga-0.02V-carga+0.026V50-carga-0.023Penet-carga-0.058PIE-carga+0.022N2b-carga-0.011SatCarga-0.014ResCarga+0.004BCMI-carga
0.91879	0.06563	0.78401	0.449S/C-0.42C3-0.617Trect+0.016D-carga+0.056CCR-carga-0.098Ni-carga+0.209V-carga+0.29 V50-carga+0.018Penet-carga+0.104PIE-carga-0.275N2b-carga+0.106SatCarga-0.028ResCarga-0.05BCMI-carga
0.79727	0.05695	0.84096	0.242S/C-0.172C3-0.236Trect-0.187D-carga-0.252CCR-carga+0.232Ni-carga+0.004V-carga-0.181V50-carga-0.151Penet-carga-0.361PIE-carga+0.711N2b-carga-0.015SatCarga+0.074ResCarga-0.08BCMI-carga
0.58772	0.04198	0.88294	-0.051S/C+0.186C3+0.14 Trect-0.233D-carga-0.14CCR-carga+0.088Ni-carga-0.09V-carga+0.795V50-carga-0.189Penet-carga-0.32PIE-carga-0.07N2b-carga+0.06 SatCarga+0.149ResCarga-0.228BCMI-carga
0.57448	0.04103	0.92397	-0.062S/C+0.334C3+0.008Trect-0.126D-carga-0.157CCR-carga+0.136Ni-carga+0.781V-carga-0.075V50-carga-0.021Penet-carga+0.24 PIE-carga+0.042N2b-carga+0.29 SatCarga+0.003ResCarga-0.256BCMI-carga
0.35272	0.02519	0.94917	0.014S/C-0.082C3+0.004Trect-0.062D-carga+0.117CCR-carga+0.259Ni-carga+0.07 V-carga-0.382V50-carga-0.438Penet-carga-0.301PIE-carga-0.489N2b-carga-0.106SatCarga+0.455ResCarga-0.131BCMI-carga
0.27956	0.01997	0.96914	-0.005S/C+0.123C3-0.025Trect+0.358D-carga-0.035CCR-carga-0.468Ni-carga+0.264V-carga+0.03 V50-carga+0.199Penet-carga-0.44PIE-carga+0.076N2b-carga+0.16 SatCarga+0.444ResCarga+0.323BCMI-carga

Eigenvectors

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	
0.0016	0.0141	0.2712	**-0.8119**	0.4486	0.2421	-0.0515	-0.0619	0.014	-0.0053	S/C
0.0315	0.0746	**0.783**	-0.061	-0.4202	-0.172	0.1862	0.3341	-0.0819	0.1229	C3
0.0201	-0.0919	-0.4529	**-0.5737**	-0.6169	-0.2364	0.1397	0.0085	0.0039	-0.0248	Trect
0.3855	-0.0871	0.0221	-0.0138	0.0164	-0.1868	-0.2334	-0.1262	-0.0623	0.358	D-carga
0.3721	0.0235	0.0425	-0.0178	0.056	-0.2524	-0.1399	-0.157	0.1171	-0.035	CCR-carga
0.3232	-0.2831	0.0716	0.0392	-0.0984	0.2319	0.0875	0.1361	0.2593	-0.4677	Ni-carga
0.2841	-0.1374	-0.2433	-0.0204	0.2088	0.0042	-0.0899	0.7809	0.0696	0.2639	V-carga
0.2791	0.0539	-0.0955	0.0262	0.2896	-0.1805	0.7954	-0.0754	-0.3822	0.0297	V50-carga
-0.3658	-0.0892	-0.0099	-0.0225	0.0178	-0.1507	-0.1895	-0.0208	-0.438	0.1986	Penet-carga
0.112	**0.5445**	-0.0582	-0.0585	0.104	-0.3605	-0.3198	0.2404	-0.3008	-0.4399	PIE-carga
0.2276	**0.2369**	-0.0929	0.0222	-0.2748	0.7114	-0.0702	0.0422	-0.4893	0.0763	N2b-carga
-0.3431	-0.2796	-0.0695	-0.0115	0.1056	-0.0153	0.0596	0.2899	-0.106	0.1601	SatCarga
-0.0398	**0.6242**	-0.0966	-0.0136	-0.0282	0.0736	0.1491	0.0028	0.4553	0.4439	ResCarga
0.3619	-0.2178	0.0792	0.0036	-0.05	-0.0801	-0.2278	-0.2565	-0.1308	0.3226	BCMI-carga

EL DETERMINANTE DE LA MATRIZ ES:

1.0123891997701504E-5

COEFICIENTE DE ESFERICIDAD DE BARTLETT

2288.621864273835

KMO

0.7638197391725657

MATRIZ ANTI-IMAGEN DE CORRELACIONES

S/C	C3	Trect	D-carga	CCR-carga	Ni-carga	V-carga	V50-carga	Penet-carga	PIE-carga	N2b-carga	SatCarga	ResCarga	BCMI-carga
0.369(a)	-0.074	-0.04	0.029	-0.0050	0.023	-0.019	-0.013	0.023	-0.024	0.027	-0.0070	0.0010	-0.028
-0.074	**0.43(a)**	0.084	-0.01	0.0010	-0.151	0.209	0.091	0.046	-0.108	0.055	-0.027	-0.023	-0.01
-0.04	0.084	**0.61(a)**	-0.053	0.0060	-0.0050	-0.01	0.043	-0.027	0.036	-0.042	0.0020	0.022	0.04
0.029	-0.01	-0.053	**0.738(a)**	-0.444	0.036	-0.438	-0.062	-0.079	-0.317	0.19	-0.024	-0.318	-0.896
-0.0050	0.0010	0.0060	-0.444	**0.854(a)**	0.109	0.155	-0.034	0.478	-0.12	0.124	0.104	0.276	0.227
0.023	-0.151	-0.0050	0.036	0.109	**0.787(a)**	-0.407	0.023	0.285	0.333	-0.042	0.439	0.448	-0.015
-0.019	0.209	-0.01	-0.438	0.155	-0.407	**0.73(a)**	-0.093	0.185	-0.205	-0.139	-0.41	-0.0010	0.281
-0.013	0.091	0.043	-0.062	-0.034	0.023	-0.093	**0.956(a)**	0.185	0.027	0.1	0.118	0.036	0.056
0.023	0.046	-0.027	-0.079	0.478	0.285	0.185	0.185	**0.867(a)**	-0.167	0.231	-0.142	0.238	0.0020
-0.024	-0.108	0.036	-0.317	-0.12	0.333	-0.205	0.027	-0.167	**0.539(a)**	-0.096	0.48	0.079	0.445
0.027	0.055	-0.042	0.19	0.124	-0.042	-0.139	0.1	0.231	-0.096	**0.849(a)**	0.179	-0.149	-0.206
-0.0070	-0.027	0.0020	-0.024	0.104	0.439	-0.41	0.118	-0.142	0.48	0.179	**0.768(a)**	0.515	0.255
0.0010	-0.023	0.022	-0.318	0.276	0.448	-0.0010	0.036	0.238	0.079	-0.149	0.515	**0.462(a)**	0.472
-0.028	-0.01	0.04	-0.896	0.227	-0.015	0.281	0.056	0.0020	0.445	-0.206	0.255	0.472	**0.713(a)**

Se obtienen cuatro factores V1 (D-carga, CCR-carga, Ni-carga, V-carga, V50-Carga, Penet-carga, Sat-carga y BCMI-carga), V2 (PIE-carga, N2b-carga, Res-carga), V3 (C3), V4 (S/C y trect) y es este conjunto de variables el que se usará para continuar con el análisis.

Tabla 9. Variables generadas por el análisis de componentes

Variables	Vector o factor
D-carga, CCR-carga, Ni-carga, V-carga, V50-Carga, Penet-carga, Sat-carga y BCMI-carga	V1 ó V1-PC1
PIE-carga, N2b-carga, Res-carga	V2 ó V2-PC1
C3	V3 ó V3-PC1
S/C y trect	V4 ó V4-PC1

Fuente: autores del proyecto

La segunda técnica es Análisis de Regresión, para esto se hará un análisis para las variables usadas (D-carga, CCR-carga, Ni-carga, V-carga, V50-Carga, Penet-carga, Sat-carga, BCMI-carga, PIE-carga, N2b-carga, Res-carga, C3 y S/C y trect) y las generadas (V1-PC1, V2-PC1, V3-PC1 y V4-PC1).

Para el primer grupo se obtuvo el siguiente resultado

```
Linear Regression Model
-----------------------

RendvDMO = 1.8444 * S/C -  0.8631 * C3 -  0.3523 * Trect -  309.9104 * D-carga -  0.4898 * CCR-carga -
0.0897 * Ni-carga -  0.0299 * V-carga -  0.0215 * PIE-   carga + 0.6181 * SatCarga + 0.5708 * BCMI-carga +
390.2287
```

Al evaluar el modelo remplazando las variables por los valores de la primera línea del *Dataset* se encontró que:

RendvDMO = 70.92885795323599

```
=== Summary ===
Correlation coefficient                  0.8804
Mean absolute error                      2.5841
Root mean squared error                  4.1729
Relative absolute error                 37.7864 %
Root relative squared error             47.4223 %
Total Number of Instances              205
```

Con estos resultados se concluye que rendvDMO es de 70.928 y depende en gran medida de D-carga, seguida de S/C y PIE-carga no influye mucho. Los valores para obtener este resultado son (S/C=6.5, C3=4.4217, Trect=115, D-carga=0.9848, CCR-carga=10.9, Ni-carga=31.8, V-carga=35, PIE-carga= 0.146, Sat-carga=50.7, y BCMI-carga= 4.24).

Se observa que Penet-carga, N2b-carga, Res-carga, V50-Carga no hacen parte del modelo lineal. El coeficiente de correlación es de 0.8804 indicando buena aceptación y los errores son bajos. Y para las generadas por el Análisis de Componentes Principales,

```
Linear Regression Model
-----------------------------
RendvDMO = -  17.2582 * V1-PC1 -  3.254* V2-PC1 + 1.7119 * V3-PC1 + 69.7247
```

Al evaluar el modelo remplazando las variables por los valores de la primera línea del *Dataset* se encontró que:

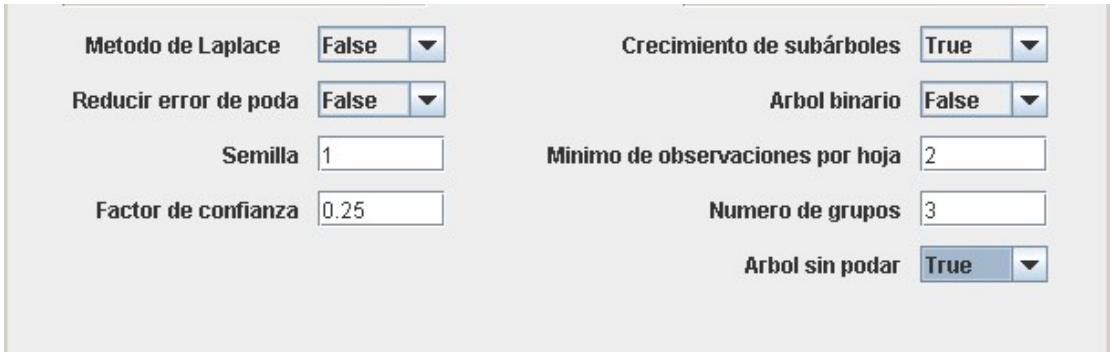
RendvDMO = 72.62600620711419

```
=== Summary ===
Correlation coefficient                  0.7482
Mean absolute error                      4.1385
Root mean squared error                  5.838
```

```
Relative absolute error               60.5161 %
Root relative squared error           66.3455 %
Total Number of Instances             205
```

Se observa que la variable rendVDMO mayor con 72.62 y V1-PC1 determina en gran medida seguida de V2-PC2. El valor de las variables son (V1-PC1= -0.2351, V2-PC1=0.1765, V3-PC1=-0.3399, V4-PC1=-0.8175) y el coeficiente de correlación es de 0.7482, indicando buena aceptación. Como los resultados cumplen satisfactoriamente los requisitos de la técnica, se aplica Árboles de Decisión. Las opciones se ven en la Figura 39 para ambos casos.

Figura 39. Opciones árbol de decisión



Fuente: Sistema de Predicción de Propiedades, SPP 2.0

Y los resultados para las variables D-carga, CCR-carga, Ni-carga, V-carga, V50-Carga, Penet-carga, Sat-carga, BCMI-carga, PIE-carga, N2b-carga, Res-carga, C3 y S/C y trect son:

```
J48 unpruned tree
------------------
Penet-carga <= 38.9
|   Trect <= 100
|   |   V-carga <= 235.38
|   |   |   C3 <= 4.4217
|   |   |   |   S/C <= 5: Medio (2.0)
|   |   |   |   S/C > 5: Alto (4.0)
|   |   |   C3 > 4.4217: Bajo (6.0/1.0)
|   |   V-carga > 235.38: Bajo (9.0)
|   Trect > 100
|   |   C3 <= 4.4217
|   |   |   Trect <= 115
|   |   |   |   S/C <= 6.5
|   |   |   |   |   ResCarga <= 26.1
|   |   |   |   |   |   S/C <= 5: Bajo (2.0)
|   |   |   |   |   |   S/C > 5: Medio (3.0/1.0)
|   |   |   |   |   ResCarga > 26.1: Bajo (23.0)
|   |   |   |   S/C > 6.5
```

```
|  |  |  |  |  PIE-carga <= 901: Medio (4.0/1.0)
|  |  |  |  |  PIE-carga > 901: Bajo (4.0)
|  |  |  Trect > 115: Bajo (15.0)
|  |  C3 > 4.4217: Bajo (22.0)
Penet-carga > 38.9
|  Trect <= 100: Alto (33.0/1.0)
|  Trect > 100
|  |  S/C <= 6.5
|  |  |  S/C <= 5
|  |  |  |  Trect <= 115
|  |  |  |  |  SatCarga <= 18.2
|  |  |  |  |  |  C3 <= 4.4217
|  |  |  |  |  |  |  N2b-carga <= 0.187: Bajo (5.0/1.0)
|  |  |  |  |  |  |  N2b-carga > 0.187: Alto (4.0)
|  |  |  |  |  |  C3 > 4.4217: Bajo (3.0/1.0)
|  |  |  |  |  SatCarga > 18.2: Medio (2.0)
|  |  |  |  Trect > 115
|  |  |  |  |  Penet-carga <= 82.9: Bajo (5.0)
|  |  |  |  |  Penet-carga > 82.9: Medio (5.0/1.0)
|  |  |  S/C > 5
|  |  |  |  C3 <= 4.4217
|  |  |  |  |  Trect <= 115
|  |  |  |  |  |  Penet-carga <= 66.6
|  |  |  |  |  |  |  Ni-carga <= 119.6: Medio (2.0)
|  |  |  |  |  |  |  Ni-carga > 119.6: Alto (2.0)
|  |  |  |  |  |  Penet-carga > 66.6: Alto (10.0)
|  |  |  |  |  Trect > 115
|  |  |  |  |  |  Ni-carga <= 103.43: Alto (3.0)
|  |  |  |  |  |  Ni-carga > 103.43: Medio (6.0)
|  |  |  |  C3 > 4.4217
|  |  |  |  |  N2b-carga <= 0.188: Medio (4.0/1.0)
|  |  |  |  |  N2b-carga > 0.188: Bajo (2.0)
|  |  S/C > 6.5
|  |  |  C3 <= 4.4217: Alto (19.0)
|  |  |  C3 > 4.4217
|  |  |  |  N2b-carga <= 0.188: Alto (4.0)
|  |  |  |  N2b-carga > 0.188: Bajo (2.0)

Number of Leaves  :    28
Size of the tree : 55

=== Summary ===

Correctly Classified Instances         197               96.0976 %
Incorrectly Classified Instances         8                3.9024 %
Kappa statistic                          0.9355
Mean absolute error                      0.0406
Root mean squared error                  0.1424
Root relative squared error             31.7076 %
Total Number of Instances              205

=== Confusion Matrix ===

  a  b  c   <-- classified as
 78  0  2 |  a = Alto
  0 95  2 |  b = Bajo
  1  3 24 |  c = Medio
```

Se debe tener en cuenta el Estadístico Kappa en la interpretación de resultados

Kappa	grado de acuerdo
< 0,00	sin acuerdo
>0,00 - 0,20	insignificante
0,21 - 0,40	discreto
>0,41 - 0,60	moderado
0,61 - 0,80	sustancial
0,81 - 1,00	casi perfecto

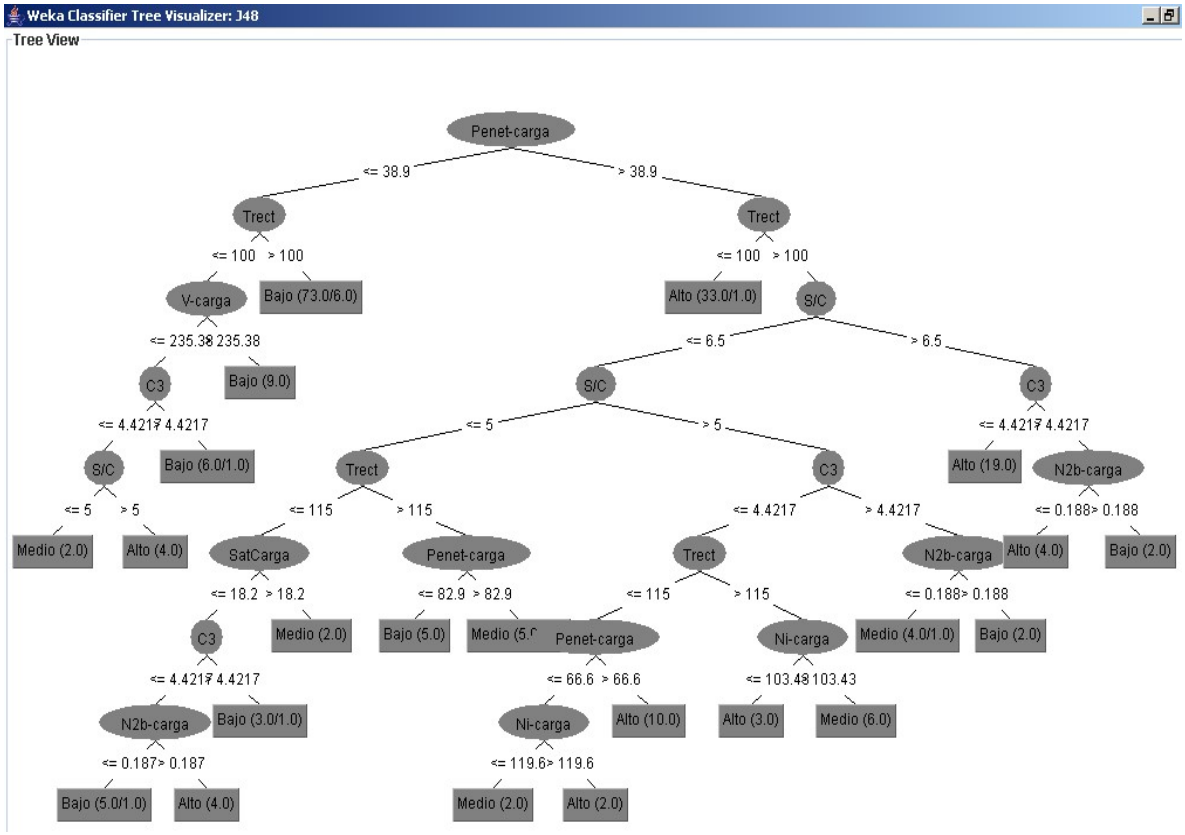
Las variables que no entran en el análisis de árbol son D-carga, CCR-carga, V50-carga y BCMI-carga.

El número de hojas del árbol es de 28 y el número de nodos 55. Las instancias clasificadas son de 197 e incorrectas 8. Se puede observar con el Estadístico Kappa con un 0.9355 con una aceptación casi perfecta. Los errores son mínimos. De las 197 hay 78 variables clasificadas en la categoría alto, 95 en bajo y 24 en medio, para un total de 197 registros. Hay varios caminos a seguir para garantizar que el DMO sea alto, (observar el Árbol Gráfico).

- Penet-carga <=38.9, trect <= 100, V-carga<=235.38, C3<=4.421, S/C>=5
- Penet-carga >38.9, trect <= 100
- Penet-carga >38.9, trect > 100, S/C<=6.5, S/C<=5, trect<=115, Sat-carga<=18.2, C3<=4.421, N2b-carga>0.187
- Penet-carga >38.9, trect > 100, S/C<=6.5, S/C>5, C3<=4.421, trect<=115, penet-carga>66.6
- Penet-carga >38.9, trect > 100, S/C<=6.5, S/C>5, C3<=4.421, trect<=115, Penet-carga<=66.6, Ni-carga>119.6
- Penet-carga >38.9, trect > 100, S/C<=6.5, S/C>5, C3<=4.421, trect>115 Ni-carga<=103.43
- Penet-carga >38.9, trect > 100, S/C>6.5, C3<=4.421
- Penet-carga >38.9, trect > 100, S/C>6.5, C3>4.421, N2b-carga<=0.188

El Árbol Gráfico se ve en la Figura 40.

Figura 40. Árbol gráfico



Fuente: Sistema de Predicción de Propiedades, SPP 2.0

Con las variables generadas por el análisis de componentes principales (V1-PC1, V2-PC2, V3-PC1, V4-PC1)

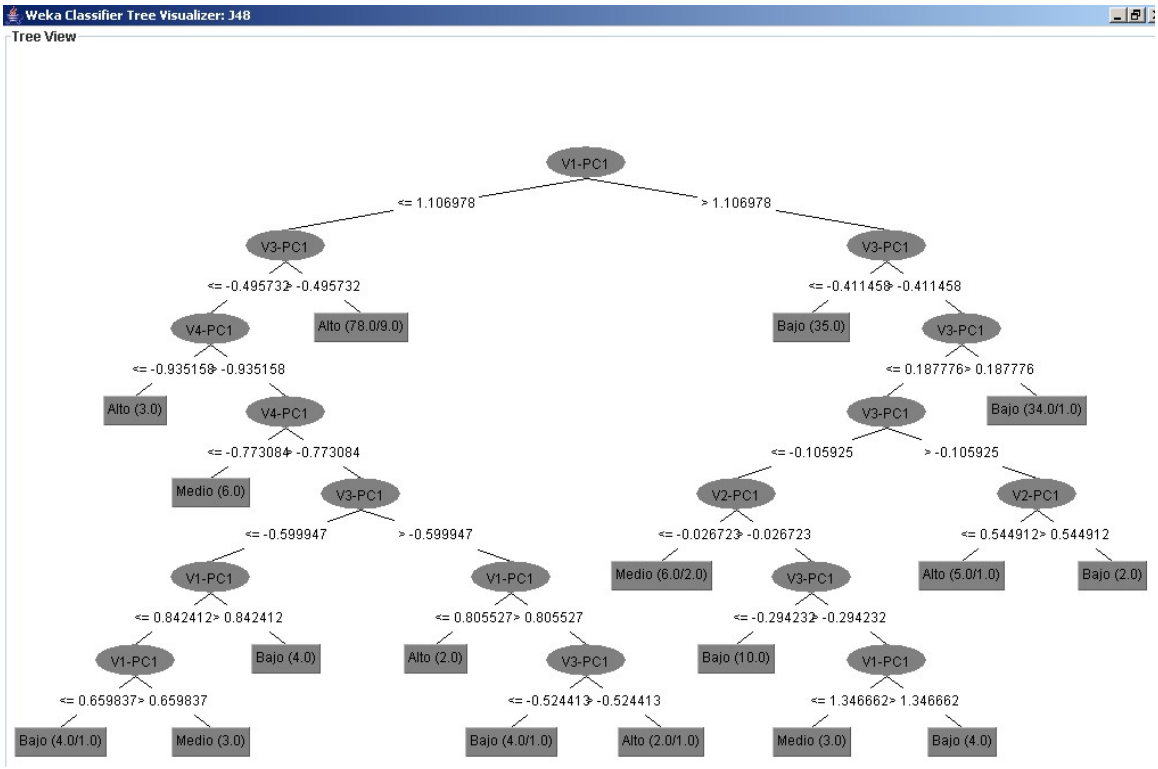
```
J48 unpruned tree
------------------
V1-PC1 <= 1.106978
|   V3-PC1 <= -0.495732
|   |   V4-PC1 <= -0.935158: Alto (3.0)
|   |   V4-PC1 > -0.935158
|   |   |   V4-PC1 <= -0.773084: Medio (6.0)
|   |   |   V4-PC1 > -0.773084
|   |   |   |   V3-PC1 <= -0.599947
|   |   |   |   |   V1-PC1 <= 0.842412
|   |   |   |   |   |   V1-PC1 <= 0.659837: Bajo (4.0/1.0)
|   |   |   |   |   |   V1-PC1 > 0.659837: Medio (3.0)
|   |   |   |   |   V1-PC1 > 0.842412: Bajo (4.0)
|   |   |   |   V3-PC1 > -0.599947
|   |   |   |   |   V1-PC1 <= 0.805527: Alto (2.0)
|   |   |   |   |   V1-PC1 > 0.805527
|   |   |   |   |   |   V3-PC1 <= -0.524413: Bajo (4.0/1.0)
|   |   |   |   |   |   V3-PC1 > -0.524413: Alto (2.0/1.0)
```

```
|   V3-PC1 > -0.495732
|   |   V3-PC1 <= 0.072917: Alto (57.0/2.0)
|   |   V3-PC1 > 0.072917
|   |   |   V3-PC1 <= 0.444218: Medio (9.0/4.0)
|   |   |   V3-PC1 > 0.444218: Alto (12.0/1.0)
V1-PC1 > 1.106978
|   V3-PC1 <= -0.411458: Bajo (35.0)
|   V3-PC1 > -0.411458
|   |   V3-PC1 <= 0.187776
|   |   |   V3-PC1 <= -0.105925
|   |   |   |   V2-PC1 <= -0.026723: Medio (6.0/2.0)
|   |   |   |   V2-PC1 > -0.026723
|   |   |   |   |   V3-PC1 <= -0.294232: Bajo (10.0)
|   |   |   |   |   V3-PC1 > -0.294232
|   |   |   |   |   |   V1-PC1 <= 1.346662: Medio (3.0)
|   |   |   |   |   |   V1-PC1 > 1.346662: Bajo (4.0)
|   |   |   V3-PC1 > -0.105925
|   |   |   |   V2-PC1 <= 0.544912: Alto (5.0/1.0)
|   |   |   |   V2-PC1 > 0.544912: Bajo (2.0)
|   |   V3-PC1 > 0.187776: Bajo (34.0/1.0)

Number of Leaves  :    19
Size of the tree : 37

=== Summary ===
Correctly Classified Instances         191               93.1707 %
Incorrectly Classified Instances        14                6.8293 %
Kappa statistic                          0.8869
Mean absolute error                      0.0694
Root mean squared error                  0.1863
Root relative squared error             41.4794 %
Total Number of Instances              205

=== Confusion Matrix ===

  a  b  c   <-- classified as
 76  0  4 |  a = Alto
  1 94  2 |  b = Bajo
  4  3 21 |  c = Medio
```

Fueron clasificadas 191 instancias correctamente y 14 incorrectamente. Los errores son bajos y Estadístico Kappa da una aceptación casi perfecta. El árbol tiene 19 hojas y 37 nodos. De las 191 hay 76 instancias clasificadas en alto, 94 en bajo y 21 en medio. Los caminos para que el DMO sea alto son:

- V1-PC1<=1.1069, V3-Pc1>-0.495732
- V1-PC1<=1.1069, V3-Pc1<=-0.495732, V4-PC1<=-0.935158
- V1-PC1<=1.1069, V3-Pc1<=-0.495732, V4-PC1>-0.935158, V4-PC1>0.773084, V3-PC1>0.599947, V1-PC1<=0.805527
- V1-PC1<=1.1069, V3-Pc1<=-0.495732, V4-PC1>-0.935158, V4-PC1>0.773084, V3-PC1>0.599947, V1-PC1>0.805527, V3-PC1>-0.524413
- V1-PC1>1.1069, V3-PC1>-0.411458, V3-PC1<=0187776, V3-PC1>-0.105925, V2-PC1<=0.544912

El Árbol Gráfico se puede apreciar a continuación:



Como la técnica de árbol de decisión dio correcta se continúa con Redes Bayesianas. Esta técnica es un mecanismo adicional de validación para asegurarse de los resultados, es decir, no es un método exclusivamente concluyente para la metodología.

Los resultados para las variables: D-carga, CCR-carga, Ni-carga, V-carga, V50-Carga, Penet-carga, Sat-carga, BCMI-carga, PIE-carga, N2b-carga, Res-carga, C3 y S/C y trect son

```
Bayes Network Classifier
------------------------

not using ADTree
#attributes=15 #classindex=14
Network structure (nodes followed by parents)

S/C(1): CatRendDMO
C3(1): CatRendDMO
Trect(2): CatRendDMO
D-carga(2): CatRendDMO
CCR-carga(2): CatRendDMO
Ni-carga(3): CatRendDMO
V-carga(2): CatRendDMO
V50-carga(3): CatRendDMO
```

```
Penet-carga(2): CatRendDMO
PIE-carga(2): CatRendDMO
N2b-carga(3): CatRendDMO
SatCarga(2): CatRendDMO
ResCarga(1): CatRendDMO
BCMI-carga(2): CatRendDMO
CatRendDMO(3):

Variables ordenadas por cardinalidad:

        Ni-carga        3
        V50-carga       3
        N2b-carga       3
        D-carga         2
        CCR-carga       2
        V-carga         2
        Penet-carga     2
        PIE-carga       2
        Trect           2
        SatCarga        2
        BCMI-carga      2
        C3              1
        ResCarga        1
        S/C             1

=== Summary ===

Correctly Classified Instances         150               73.1707 %
Incorrectly Classified Instances        55               26.8293 %
Kappa statistic                          0.5474
Mean absolute error                      0.1771
Root mean squared error                  0.3872
Root relative squared error             86.2084 %
Total Number of Instances              205

=== Confusion Matrix ===

  a  b  c   <-- classified as
 67  5  8 |  a = Alto
  9 81  7 |  b = Bajo
 18  8  2 |  c = Medio
```

Las variables con más probabilidad de influir sobre el DMO V50-carga, Ni-carga y N2b-carga. Fueron clasificadas 150 instancias correctamente y 55 incorrectamente. El Estadístico Kappa da una aceptación moderada y los errores dan bajo. Se clasificaron 67 instancias en alto, 91 en bajo y 2 en medio de las 150. Para las variables generadas por el análisis de componentes principales V1-PC1, V2-PC1, V3-PC1 y V4-PC1 los resultados son

```
Bayes Network Classifier
-----------------------

not using ADTree
#attributes=5 #classindex=4
Network structure (nodes followed by parents)

V1-PC1(2): CatRendDMO
```

```
V2-PC1(1): CatRendDMO
V3-PC1(3): CatRendDMO
V4-PC1(1): CatRendDMO
CatRendDMO(3):

Variables ordenadas por cardinalidad:

        V3-PC1          3
        V1-PC1          2
        V2-PC1          1
        V4-PC1          1

=== Summary ===

Correctly Classified Instances         167               81.4634 %
Incorrectly Classified Instances        38               18.5366 %
Kappa statistic                          0.6933
Mean absolute error                      0.2071
Root mean squared error                  0.3055
Root relative squared error             68.0293 %
Total Number of Instances              205


=== Confusion Matrix ===

  a  b  c   <-- classified as
 69  5  6 |  a = Alto
  1 86 10 |  b = Bajo
  8  8 12 |  c = Medio
```

Las variables con más probabilidad de influir sobre el DMO son V3-PC1. Fueron clasificadas 167 instancias correctamente y 38 incorrectamente. El estadístico kappa da una aceptación sustancial (mejor que el anterior) y los errores dan bajo. Se clasificaron 69 instancias en alto, 86 en bajo y 12 en medio de las 167.

3.7 SOLUCIÓN

En esta etapa el usuario podrá concluir y tomar decisiones o realizar nuevos análisis, (no necesariamente ligados a la serie de pasos) y comparar entre ellos el mejor. Por ejemplo, el usuario puede dividir la base de datos en la mitad o en lo que considere necesario y hacer un análisis de estos nuevos *datawarehouse*; o incluir nuevas categorías como (bajo-bajo, bajo, medio, alto, alto-alto), y así tendrá más bases para determinar las variables independientes que hacen más óptimo en DMO.

Para el modelo A puede escoger cualquiera de los conjuntos dados ya que como se pudo observar a través de las pruebas dan buenos resultados. Para el modelo B sólo puede observar como se agrupan las diferentes variables y determinar cuál conglomerado da una mejor aproximación en el análisis de regresión.

4. CONCLUSIONES

La primera técnica en ser aplicada por el usuario es el Análisis de Componentes Principales ya que permite eliminar variables no correlacionadas con la población y agruparlas en factores que facilitará a través de las otras técnicas concluir cuáles son las variables que satisfacen los objetivos planteados por el usuario.

Las nuevas funcionalidades que se incluyeron en el análisis de componentes principales como el determinante de la Matriz de Correlación debe acercarse a cero, pero sin serlo; el Test de esfericidad prueba la hipótesis nula que afirma que las variables no están correlacionadas en la población; los valores en el Índice Kaiser-Meyer-Olkin debe ser mayor a 0.7 y la matriz anti-imagen de correlaciones parciales muestra en los valores de la diagonal donde se ubica la media de adecuación muestral de cada variable (MSA) es muy bajo (< 0.3) se debe descartar esa variable y se debe realizar de nuevo el análisis de componentes principales.

El árbol mostrado gráficamente en una nueva ventana le permite al usuario interpretar mejor cuáles son las reglas que debe seguir para la toma de decisiones generadas por el algoritmo J48 en la técnica Árboles de Decisión.

Al modificar el formato de presentación de la Ecuación del análisis de regresión se hace una interpretación más clara de cómo está constituida la variable dependiente respecto de las independientes.

Al incluir el menú de ayuda al prototipo se hizo un aporte significativo ya que el usuario cuenta con un soporte sobre el uso adecuado para evitar excepciones en el prototipo y entender la interpretación de los valores a ingresar en las diferentes técnicas.

La técnica de *clustering* solo permite determinar cómo están agrupadas las variables y en que medida afectan o contribuyen en el rendimiento del DMO.

Las variables C3, i-C4 y n-C4 no pueden ir juntas en el Análisis de Componentes Principales porque ser el dato constante la varianza es cero afectando significativamente el los resultados.

Las variables CarARo-carga y AromCarga no son muy importantes para el usuario por eso en algunos casos no se tomaban en cuenta, ya que afectaban los resultados.

La variable S/C es muy importante para el investigador, cuando se realizaba el Análisis de Componentes Principales en muchos casos está era la única variable que no cumplía el requisito de la medida de adecuación de la muestra, MSA. Así que se debió hacer muchas clasificaciones y agrupaciones entre las variables para hacer que se aceptarán todas las condiciones de esta técnica.

La técnica Redes Bayesianas es un elemento adicional de validación ya que a través de las pruebas y análisis que se desarrollaron no se encontró gran aplicabilidad en los procesos de Minería de Datos.

La serie de pasos ha sido de gran utilidad para el usuario permitiendo realizar actividades guiadas a través del análisis, haciendo más fácil la interpretación de cada una de las técnicas de Minería de Datos presentes en el prototipo.

El trabajo de investigación ha sido muy valioso y enriquecedor ya que se integran áreas como ingeniería de software, programación, estadística y matemáticas logrando afianzar conocimientos que nos permitan desarrollarnos íntegramente en organizaciones.

BIBLIOGRAFÍA

AERTIA SOFTWARE, se puede ver la descripción, precios, demos y descargar Monarch [online, Articulo], 2005. [Citado 12 septiembre 2006]. Disponible en Internet: <http://www.aertia.com/productos.asp?pid=231>

HERNANDEZ GUERRA, Alejandro. Aprendizaje Automático: Árboles de Decisión. Universidad Veracruzana, México. 2004. [online, Articulo]. [Citado el 24 de febrero 2006]. Disponible en Internet: <http://www.uv.mx/aguerra/teaching/ MIA/MachineLearning/clase07.pdf> p. 6-8.

ANSWERMATH, tutoriales de Minería de Datos [online, Tutorial] 2005. [Citado el 8 de Febrero]. Disponible en Internet: <http://www.answermath.com/ mineria_de_datos.htm>

BERZAL GALIANO, Fernando y TALAVERA CUBERO. Departamento de ciencias de la computación e inteligencia artificial, ETS-Ingeniería informática, Universidad de Granada [online, Articulo]. [Citado el 20 de febrero 2006]. Disponible en Internet: <http://elvex.ugr.es/etexts/spanish/proyecto/cap5.pdf> P. 3.

BRESSÁN, Griselda. Trabajo monográfico de adscripción. Lic. En Sistemas de Información Almacenes de Datos y Minería de Datos. [Online, Articulo], 2003. [Citado el 5 de febrero de 2006] Disponible en Internet: <http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatos Bressan.htm>

CHEN and WANG. Discovery of Operational Spaces from Process Data for Production of Multiple Grades of Products Ind. Eng. Chem. Res. 2000, 39, p. 2378-2383

DAEDALUS - DATA, Decisions and Language, S. A. Minería de Datos [online, Articulo] 2006. [Citado el 24 de febrero 2066]. Disponible en Internet: <http://www.daedalus.es/AreasMD-E.php> y <http://www.daedalus.es/AreasMD Fases-E.php>

DEBASHIS, Neogi and SCHLAGS, Cory. Process Multivariate Statistical Analysis of an Emulsion Batch Ind. Eng. Chem. Res. 1998, 37, 3971-3979

Departamento de Sistemas Informáticos y Computación. Valencia España. Aprendizaje de árboles de decisión [online, Articulo]. [Citado el 22 de febrero 2006]. Disponible en Internet: <http://www.dsic.upv.es/asignaturas/facultad/apr/decision.pdf> P. 5.

Estadístico, es el sitio Web especialistas en consultoría y formación estadística, integrado por expertos en los programas SPSS, SAS, CLEMENTINE entre otros, en la Web de data Mining Institute encontrará todo lo referente a la estadística: cursos, artículos, software, enlaces, consultoría, libros, diccionario estadístico y tests. [Online, Artículo estadístico] 2004. [Citado el 07 de Febrero 2006] Disponible en <http://www.estadistico.com/arts.html?20001023> y <http://www.estadistico.com /arts.htm?20001106> [Citado el 25 de agosto de 2006] <http://www.estadistico.com/dic.html?p=4135&PHPSESSID= 83d26dfa82 897dc24a9ec5c8225dd61a>

Gams.com es el sitio Web oficial The General Algebraic Modeling System (GAMS). http://www.gams.com/docs/intro.htm

HAIR, ANDERSON; TATHAM y BLACK. Análisis Multivariante, quinta edición, Prentice Hall, 2001, p. 143-148, 347-349, 767, 779

HERNÁNDEZ ORALLO, José; RAMÍREZ QUINTANA; Maria José y FERRI RAMÍREZ, Cesar. Introducción a la Minería de Datos. 2005. Editorial Pearson, p. 266-269

HILLER y LIBERMAN Investigación de operaciones, séptima edición Junio de 2003, editorial McGraw Hill, p. 654, 664 - 669.

I-MINER 3.0 - ADDLINK Software Científico. I-Miner 3.0 Software de alto nivel para usuarios no iniciados, Minería de Datos al alcance de todos [online, Información software], 2006. [Citado 12 septiembre 2006]. Disponible en Internet: <http://www.addlink.es/ productos.asp?pid=277>.

INFLEXA, ¿Qué es Minería de Datos? [Online, Articulo] 2003 santiago de chile. [Citado el 30 de enero 2006] Disponible en Internet: <http://www.inflexa.com/jsp/template.jsp?pag=mineria-datos.htm&mnu=mnu-mineria.htm>

INSTITUTO COLOMBIANO DE NORMAS TÉCNICAS Y CERTIFICACIÓN. Compendio tesis y otros trabajos de grado. Quinta actualización. Santafé de Bogotá D.C.: ICONTEC, 2002.

Instituto de Computación, Universidad de la República, Montevideo-Uruguay. Árboles de decisión [online, Articulo]. [Citado el 22 de febrero 2006]. Disponible en Internet: < http://www.fing.edu.uy/inco/cursos/aprendaut/transp/arboles.pdf> P.8

JAECKLE, Christiane, and MACGREGOR, Jhon. Product design through multivariate statistical analysis of process data. Dept. of Chemical Engineering, AIChE Journal mayo 1998, vol. 44, No. 5.

MORENO GARCÍA, María. MIGUEL QUINTALES, Luís. GARCÍA PEÑALVO, Francisco y POLO MARTÍN, José. Aplicación de técnicas de Minería de Datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software [online, Articulo]. [Citado el 11 de febrero 2006]. Disponible en Internet: <http://www.sc.ehu.es/jiwdocoj/remis/docs/minerw.pdf>

NEOGI y SCHLags. Multivariate Statistical Analysis of an Emulsion Batch Process. Ind. Eng. Chem. Res. 1998, 37, 3971-3979

PEREA, Manuel. Associate Professor Universidad de Valencia. Bloque III. Caracterización de la relación entre variables. 2006 [online, Presentacion power point]. [Citado el 30 de agosto 2006] Disponible en Internet: <http://www.uv.es/~mperea/T8_APD.ppt>

PLA, Laura. Análisis de multivariado: método de componentes principales. Washington: Secretaría General de la Organización de los Estados Americanos, Programa Regional de Desarrollo Científico y Tecnológico. p. 1-17.1986. Serie de Matemáticas, monografía No 27.

RESAMPLING STATS, INC. es el sitio Web oficial de Resampling Stats donde se encuentra toda la información pertinente a Xlminer [online, Software], 2006. [Citado el 12 de septiembre 2006]. Disponible en Internet: <http://www.resample.com/xlminer/>

RUMBAUGHT, James; JACOBSON, Ivar y BOOCH, Grady. El Lenguaje Unificado de Modelado. Manual de Referencia. Edición año 2000. Editorial Addison Wesdey.

SANTEN. KOOT. ZULLO. Statistical data analysis of a chemical plant. 1997. Computers chem. engng. Vol 21. suppl., pp. s1123-s1129

SERVENTE, Magdalena. Algoritmos TDIDT aplicados a la Minería de Datos inteligente. 2002 [online, Articulo]. [Citado el 21 de octubre 2006]. Disponible en Internet: <http://www.fi.uba.ar/laboratorios/lsi/servente-tesisingenieria informatica.pdf > p. 77-89.

SPIEGEL, Murry y STEPHENS, Larry. Estadística. Editorial Mc Graw Hill, tercera edición.

SPSS Inc, Acerca de SPSS [online, Software], 2005. [Citado el 12 de septiembre 2006]. Disponible en Internet: <http://www.spss.com/la/>

VALLE, Sergio. WEIHUA. And QIN, Joe. Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods Ind. Eng. Chem. Res. 1999, 38, 4389-4401

WANG and LI Combining Conceptual *Clustering* and Principal Component Analysis for State Space Based Process Monitoring Ind. Eng. Chem. Res. 1999, 38, 4345-4358

WANG and MCGREAVY. Automatic Classification for Mining Process Operational Data, Ind. Eng. Chem. Res. 1998, 37, 2215-2222

Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Se pude descargar software y documentación [online, Software]. [Citado el 03 de marzo de 2006] Disponible en Internet: <http://www.cs.waikato.ac.nz/ml/*weka*/>

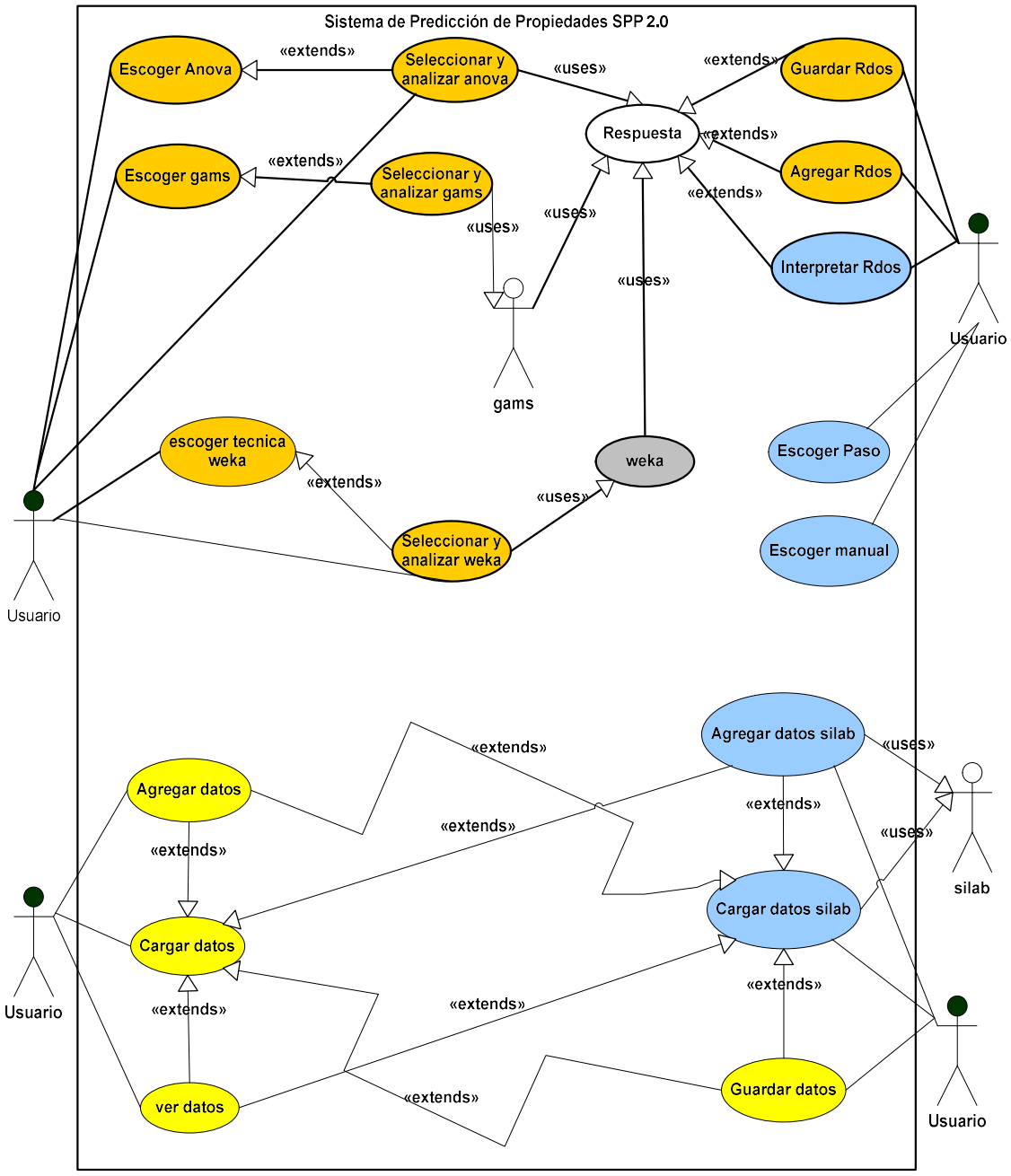
WITTEN Ian and EIBE Frank. Data Mining Practical Machine Learning Tools and Techniques. Editorial Morgan Kaufmann, second edition, 2005, p. 119-121.

WOLFF Carmen Gloria. La Tecnología Datawarehousing. 1999 [online, Articulo]. [Citado el 27 de agosto 2006]. Disponible en Internet: <http://www.inf.udec.cl/revista/ediciones/edicion3/cwolff.PDF> p. 2.

Yale. Yet Another Learning Environment? [Online, Software], 2006. [Citado el 12 de septiembre 2006]. Disponible en Internet: <http://rapid-i.com/>

ANEXOS

Anexo A. Diagrama de casos de uso



El usuario es siempre el mismo, se representa varias veces para una mejor presentacion y entendimiento para el lector.
Asi como el actor gams está por fuera del sistema.

ESPECIFICACIÓN DE LOS CASOS DE USO

Especificación Caso de Uso: Agregar Datos

Agregar Datos

Descripción
Este caso de uso es una opción de la interfaz inicio o ventana principal del menú archivo y submenú Agregar Datos/desde archivo. Es inicializado por el usuario cuando desea Agregar al *dataset* más datos desde un archivo. Finaliza una vez que se termina este proceso.

Flujo de Eventos

Flujo Básico
Inicia desde la opción Cargar Datos/Archivo de Datos del menú Archivo en la ventana principal. Esta opción abre una nueva ventana para que el usuario seleccione la ubicación donde está el archivo con extensión .csv con los datos que desea agregar.

Requisitos Especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina Java para correr la aplicación y la carpeta SPP en C:/Archivos de programa.

Pre-condiciones

< Pre-condición Uno >
Debe haberse cargado los datos en el sistema, (desde un archivo o desde el silab) lo cual se comprobará con el *Dataset* del caso de uso ver datos.

Post-condiciones

< Post-condición Uno >
Los datos son agregados al final del *dataset*.

Especificación Caso de Uso: Cargar Datos

Cargar Datos

Descripción
Este caso de uso es inicializado por el usuario cuando selecciona la opción Cargar Datos/Desde Archivo del menú Archivo para cargar los datos que va a utilizar, es decir, son los datos que serán cargados en el *dataset* para analizarlos con las técnicas seleccionadas. Finaliza una vez el usuario termina de usar la aplicación o cargue nuevos datos.

Flujo de Eventos

Flujo Básico
Inicia desde la opción Cargar Datos del menú Archivo en la ventana principal. El usuario selecciona el submenú Archivo de datos y esta abre una ventana para que el usuario indique la ubicación del archivo con extensión .csv el cual contiene los datos y estos son agregados al *dataset* que es ejecutado por la clase MainData.

Flujo Alterno
Se activa el botón ver Datos, se activa del menú Archivo las opciones Guardar Datos y Agregar Datos: Archivo de datos y Silab. Si el archivo no es correcto no se activa ninguna opción anterior

Requisitos Especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina Java para correr la aplicación y la carpeta SPP en C:/Archivos de programa.

Especificación Caso de Uso: Agregar Datos Silab

Agregar Datos silab

Descripción
Este caso de uso es inicializado por el usuario cuando desea Agregar al *dataset* más datos desde la base de datos, *Silab*. Finaliza una vez que se termina este proceso.

Flujo de Eventos

Flujo Básico
Inicia desde la opción Agregar Datos/Silab del menú Archivo en la ventana principal. Esta opción abre una nueva ventana para que el usuario seleccione un archivo con extensión .csv con las referencias de los datos, es decir, los simples id de los elementos, después a través de la clase conexión lee internamente el puerto, ip, bd, user y pass para hacer la conexión y los elementos y alias de la carga, el dmo, el demex y el solvente. Después de esto se hacen las consultas respectivas y se cargan en el *dataset*.

Requisitos especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina Java para correr la aplicación, ser trabajador el ICP y tener acceso a la base de datos, Silab y la carpeta SPP en C:/Archivos de programa.

Pre-condiciones

< Pre-condición Uno >
Haber cargado datos previamente desde un archivo o desde la base de datos SILAB.

Post-condiciones

< Post-condición Uno >
Los datos agregados se localizan al final del *dataset*, o después de los ya cargados.

Especificación Caso de Uso: Cargar Datos Silab

Cargar Datos silab

Descripción
Este caso de uso es inicializado por el usuario cuando desea Cargar al *dataset* datos desde la base de datos, silab. Finaliza una vez que se termine este proceso.

Flujo de Eventos

Flujo Básico
Inicia desde la opción Cargar Datos/Silab del menú Archivo en la ventana principal. Esta opción abre una nueva ventana para que el usuario seleccione un archivo con extensión .csv con las referencias de los datos, es decir, los *samples id* de los elementos, después a través de la clase conexión lee internamente el puerto, ip, bd, user y pass para hacer la conexión y los elementos y alias de la carga, el dmo, el demex y el solvente. Después de esto se hacen las consultas respectivas y se cargan en el *dataset*.

Flujo Alterno
Se activa el botón ver Datos, se activa del menú Archivo las opciones Guardar Datos y Agregar Datos: Archivo de datos y Silab. Si el archivo no es correcto no se activa ninguna opción anterior

Requisitos especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina java para correr la aplicación, ser trabajador el ICP y tener acceso a la base de datos, silab y la carpeta SPP en C:/Archivos de programa.

Pre-condiciones

< Pre-condición Uno >
Haber Cargado datos previamente desde un archivo o desde la base de datos SILAB.

Especificación Caso de Uso: Ver Datos

Ver Datos

Descripción
Este caso de uso es inicializado por el usuario cuando desea ver los datos que está utilizando, es decir, ver los datos que están en el *dataset*. Finaliza una vez el usuario cierra la ventana, si el usuario cambia algún valor de la variable esta será actualizada automáticamente.

Flujo de Eventos

Flujo Básico
Inicia desde la opción ver Datos representados por un botón en la ventana principal y muestra el *dataset* con las variables. Esto se hace a través de la clase Maindata que se comunica con el *dataset*.

Requisitos especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina Java para correr la aplicación y la carpeta SPP en C:/Archivos de programa.

Pre-condiciones

< Pre-condición Uno >
Haber cargado datos previamente desde un archivo o desde la base de datos SILAB.

Especificación Caso de Uso: Guardar Datos

Ver Datos

Descripción
Este caso de uso es inicializado por el usuario cuando desea guardar los datos que esta utilizando en un archivo .csv

Flujo de Eventos

Flujo Básico
Cuando se active la opción se muestra una interfaz para que el usuario guarde un archivo .csv en la ruta y con el nombre que proporcione, se hace con el método guardaDatos de la clase principal (inicio). Finaliza una vez se guarde el archivo.

Requisitos especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina Java para correr la aplicación y la carpeta SPP en C:/Archivos de programa.

Pre-condiciones

< Pre-condición Uno >
Haber cargado datos previamente desde un archivo o desde la base de datos SILAB.

Especificación Caso de Uso: Escoger Anova

Escoger Anova

Descripción
Este caso de uso es inicializado por el usuario cuando se escoge la opción análisis de varianza/anova del menú técnicas y oprime el botón analizar que se activa y finaliza cuando se muestran los resultados en la interfaz.

Flujo de Eventos

Flujo Básico
Inicia desde la opción seleccionada del menú Técnicas en la ventana principal y cuando el usuario seleccione la opción o el botón analizar. Se llama a la clase SeleccionarVariables donde el usuario escoge la variable dependiente y las variables independientes.

Flujo Alterno
Se activa el botón analizar y se llama a la clase Anova donde se calcula las medias por fila y columna, la media de medias y calcula la suma total de cuadrados (STT), suma de cuadrados del tratamiento – intermuestral (SSTR), Suma de cuadrados del error o intramuestral (SEE), después se accede a la clase AnovaCalculos para comparar con el tstudent y determinar cuales variables son o no significativas.

Requisitos especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina Java para correr la aplicación y la carpeta SPP en C:/Archivos de programa.

Pre-condiciones

< Pre-condición Uno >
Haber Cargado datos previamente desde un archivo o desde la base de datos SILAB.

Post-condiciones

< Post-condición Uno >
Extiende el caso de uso respuesta que es controlado con la clase Formrespuesta.

Especificación Caso de Uso: Escoger Gams

Escoger Gams

Descripción
Este caso de uso es inicializado por el usuario cuando en la ventana principal del prototipo escoge Gams, después se abre una ventana para que se seleccione variables y finaliza una vez se determinen los resultados.

Flujo de Eventos

Flujo Básico
Inicia desde la opción seleccionada del menú Técnicas en la ventana principal y cuando el usuario seleccione la opción o el botón analizar. Se llama a la clase SeleccionarVariables donde el usuario escoge la variable dependiente y las variables independientes.

Flujo Alterno
Se activa el botón analizar y se llama a la clase gams donde se escribe en una ruta por defecto los datos y se accede a la clase llamadogams para que lea de la ruta el modelo y los datos y ejecute el método *runtime* con estos parámetros, el software gams.exe realiza la operación respectiva y retorna la respuesta en un modelo.dat localizado en la ruta por defecto.

Requisitos especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina java para correr la aplicación y la carpeta SPP en C:/Archivos de programa.

Pre-condiciones

< Pre-condición Uno >
Haber Cargado datos previamente desde un archivo o desde la base de datos SILAB.

Post-condiciones

< Post-condición Uno >
Extiende el caso de uso respuesta que es controlado con la clase Formrespuesta.

Especificación Caso de Uso: Escoger técnica *Weka*

Escoger técnica *Weka*

Descripción
Este caso de uso es inicializado por el usuario cuando escoge entre las cinco técnicas de *Weka* disponibles (Componente Principales, Cluster, Árboles de decisión, Redes Bayesianas y Análisis de regresión). Finaliza una vez el usuario termina de usar la aplicación.

Flujo de Eventos

Flujo Básico
Inicia desde la opción seleccionada del menú Técnicas en la ventana principal y cuando el usuario seleccione la opción o el botón analizar. Se llama a la clase SeleccionarVariables donde el usuario escoge la variable dependiente y las variables independientes así como las opciones propias de cada técnica.

Flujo Alterno
Si la técnica es Análisis de Componentes Principales, se activa el botón analizar y se llama a la clase principal la cual hace conexión con *Weka* y este retorna la matriz de correlación la cual es usada en la clase Componentes (creada por el autor), esta clase permite calcular el determinante de la matriz, el KMO, el Test de esfericidad y la Matriz Anti-imagen los cuales determinan cuales son las variables a eliminar y si es posible o no realizar este análisis.

Si la técnica es Árboles de Decisión se activa el botón analizar y se llama a la clase clasificador la cual hace conexión con W*eka* este retorna los resultados y después a través de nuevas líneas de código se construye el árbol gráficamente para ser interpretado más fácilmente.

Si la técnica es Análisis de Regresión se activa el botón analizar y se llama a la clase clasificador la cual hace conexión con *Weka* este retorna los resultados y después a través de nuevas líneas de código se construye la Ecuación del modelo lineal para ser interpretado más fácilmente.

Si la técnica es Redes de Bayes se activa el botón analizar y se llama a la clase clasificador la cual hace conexión con *Weka* este retorna los resultados de la técnica.

Si la técnica es *Cluster* se activa el botón analizar y se llama a la clase cluster la cual hace conexión con *Weka* este retorna los resultados de la técnica.

Requisitos especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina Java para correr la aplicación y la carpeta SPP en C:/Archivos de programa.

Pre-condiciones

< Pre-condición Uno >
Haber Cargado datos previamente desde un archivo o desde la base de datos SILAB, activar y dar clic en el botón analizar, seleccionar variables y dar clic en analizar .

Post-condiciones

< Post-condición Uno >
Una vez se obtienen los resultados son retornados a la clase SeleccionarVariables la cual a través de la clase Formrespuesta se muestran los resultados.

Especificación Caso de Uso: Respuesta/Guardar Rdos

Respuesta/GuardarRdos

Descripción
Este caso de uso es inicializado por el caso de uso analizar técnica *Weka*, analizar gams o analizar anova. Se encuentra en la interfaz de respuesta. Finaliza una vez se lleve a cabo la tarea.

Flujo de Eventos

Flujo Básico
Este caso de uso es inicializado por el usuario cuando en la interfaz de resultados escoge guardar resultados o desde el menú archivo Guardar/resultados. Los resultados son almacenados en un archivo .rtf y finaliza una vez el proceso termina. El primero en la clase Formrespuesta y el segundo en la clase inicio se ejecuta el método guardarResultadoActual el cual solicita el nombre del archivo .rtf y la ruta.

Requisitos especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina Java para correr la aplicación y la carpeta SPP en C:/Archivos de programa.

Pre-condiciones

< Pre-condición Uno >
Haber Cargado datos previamente desde un archivo o desde la base de datos SILAB y haber analizado alguna técnica.

Especificación Caso de Uso: Respuesta/Agregar Rdos

Respuesta/GuardarRdos

Descripción
Este caso de uso es inicializado por el caso de uso analizar técnica W*eka*, analizar Gams o analizar ANOVA. Se encuentra en la interfaz de respuesta. Finaliza una vez se lleve a cabo la tarea.

Flujo de Eventos

Flujo Básico
Este caso de uso es inicializado por el usuario cuando en la interfaz de resultados escoge agregar resultados. Los resultados son almacenados en un archivo por defecto resultadosglobales.rtf y finaliza una vez el proceso termina.

Requisitos especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina Java para correr la aplicación y la carpeta SPP en C:/Archivos de programa.

Pre-condiciones

< Pre-condición Uno >
Haber Cargado datos previamente desde un archivo o desde la base de datos SILAB y haber analizado alguna técnica.

Especificación Caso de Uso: Respuesta/Interpretar Rdos

Respuesta/Interpretar Rdos

Descripción
Este caso de uso es inicializado por el caso de uso analizar técnica W*eka*, analizar Gams o analizar ANOVA. Se encuentra en la interfaz de respuesta. Finaliza una vez se lleve a cabo la tarea.

Flujo de Eventos

Flujo Básico
Este caso de uso es inicializado por el usuario cuando en la interfaz de resultados escoge interpretar resultados. Se llama a la clase VentanaAyuda con el parámetro titulo de la técnica.

Requisitos especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina Java para correr la aplicación y la carpeta SPP en C:/Archivos de programa.

Pre-condiciones

< Pre-condición Uno >
Haber Cargado datos previamente desde un archivo o desde la base de datos SILAB y haber analizado alguna técnica.

Especificación Caso de Uso: Escoger paso metodología

Escoger paso metodología

Descripción
Este caso de uso es inicializado por el usuario cuando desee ver los pasos de la metodología y finalizan cuando se cierre la ventana o la aplicación.

Flujo de Eventos

Flujo Básico
Este caso de uso es inicializado por el usuario cuando desde el menú Ayuda/Metodología escoge un paso. Se llama a la clase VentanaAyuda con el parámetro nombre de la acción.

Requisitos especiales

< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina Java para correr la aplicación y la carpeta SPP en C:/Archivos de programa.

Especificación Caso de Uso: Escoger Manual

Escoger Manual

Descripción
Este caso de uso es inicializado por el usuario cuando desee ver el manual de usuario del prototipo y finaliza cuando el usuario cierre la ventana o la aplicación.

Flujo de Eventos

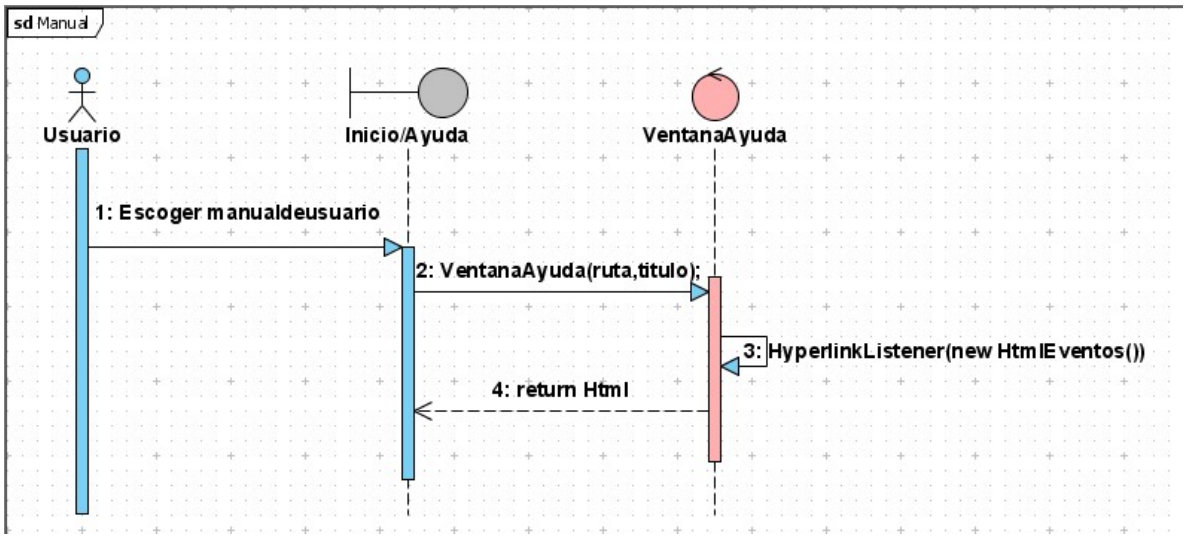
Flujo Básico
Este caso de uso es inicializado por el usuario cuando selecciona en el menú Ayuda/Manual de usuario. Se llama a la clase VentanaAyuda con el parámetro nombre de la acción.

Requisitos especiales

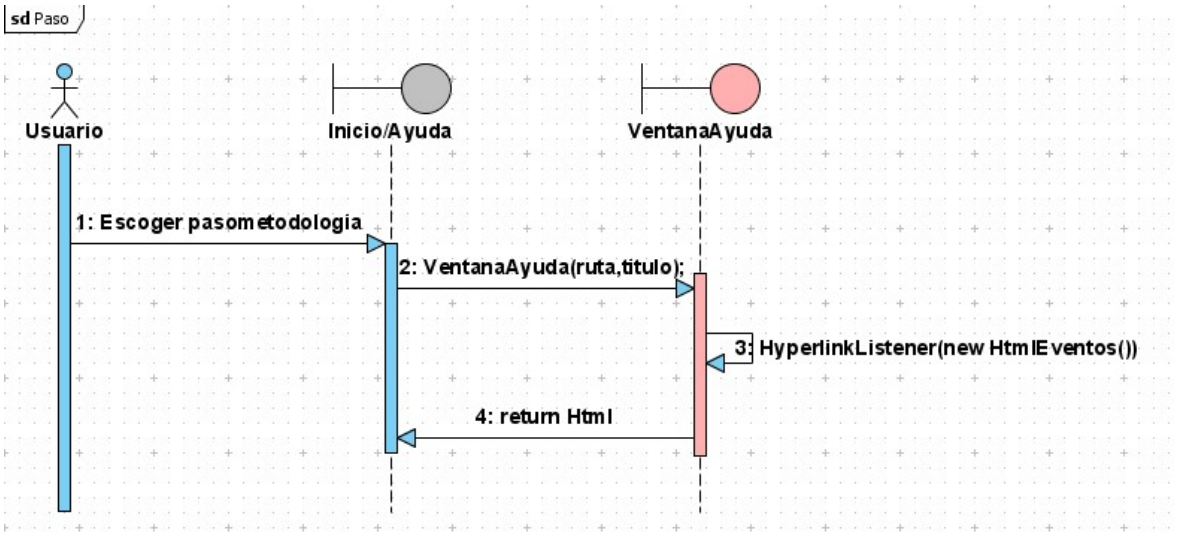
< Primer Requisito Especial >
Para el caso de los usuarios WINDOWS es necesaria la máquina Java para correr la aplicación y la carpeta SPP en C:/Archivos de programa.

Anexo B. Diagrama de secuencias

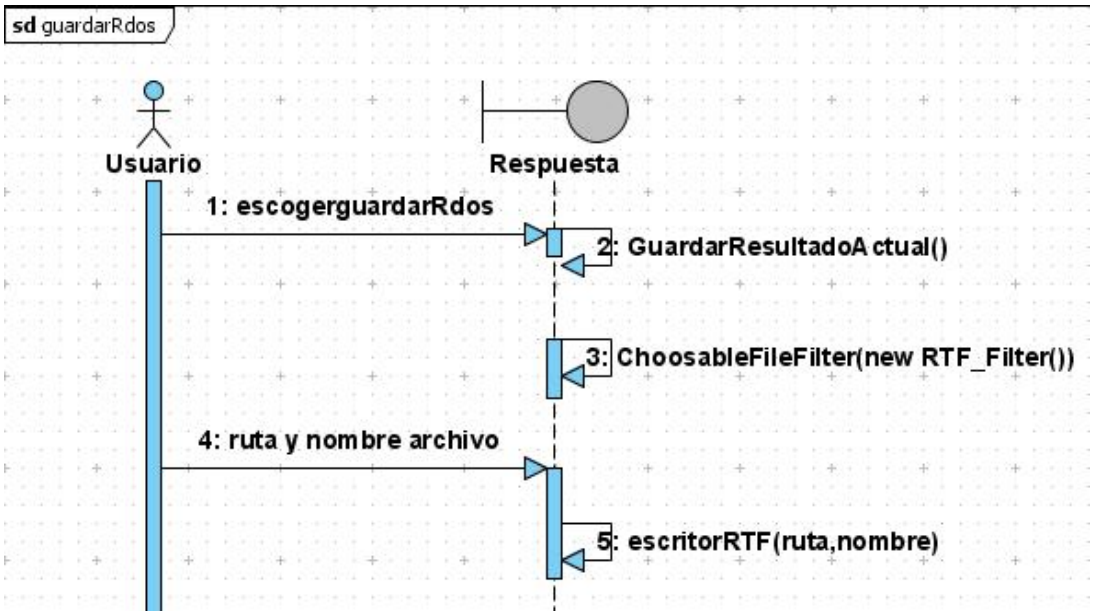
Caso de uso, escoger manual de usuario



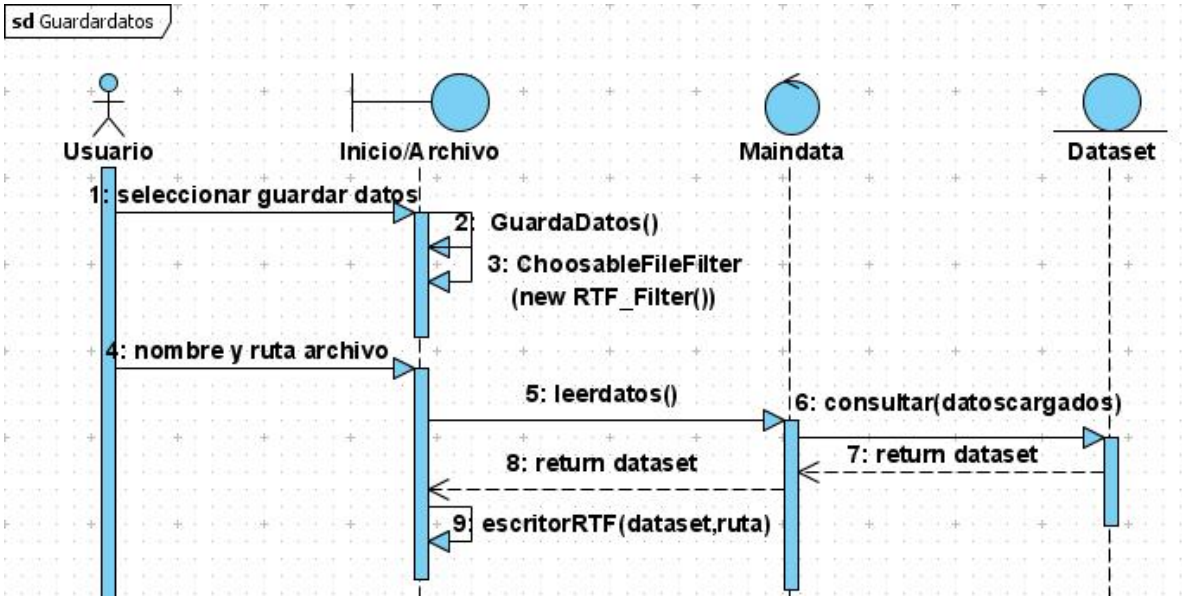
Caso de uso, escoger paso metodología



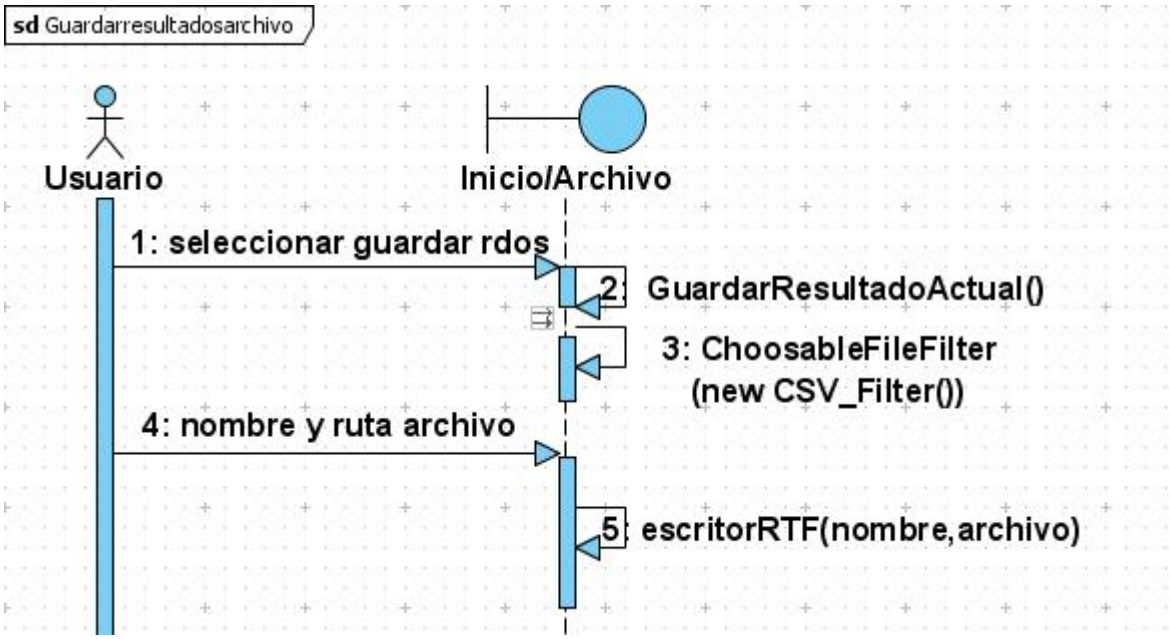
Caso de uso, Guardar resultados



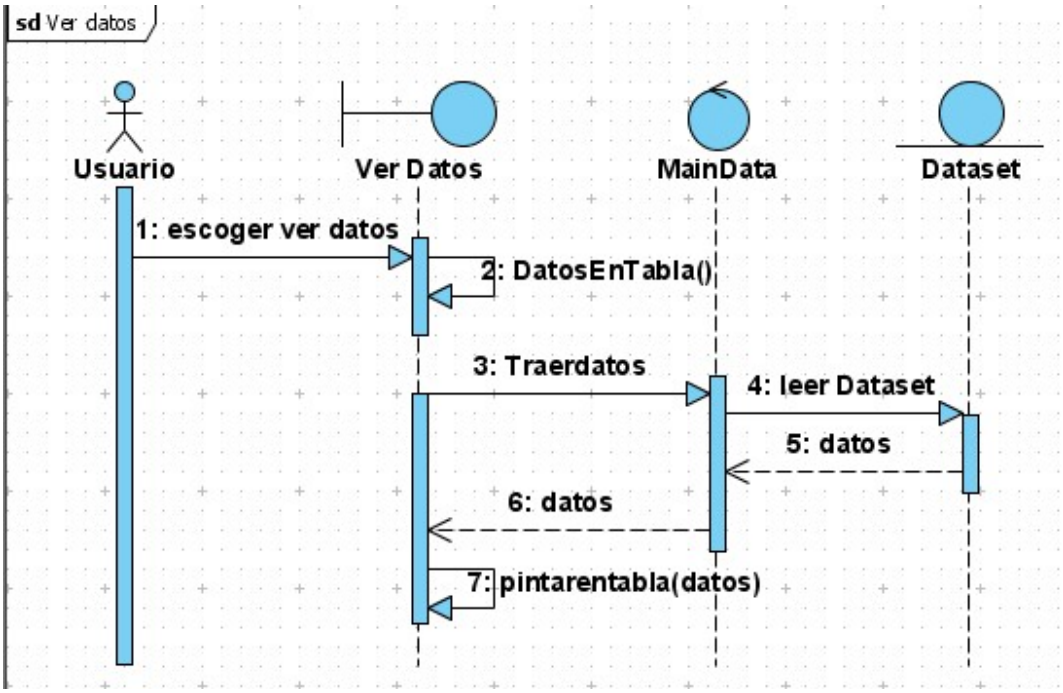
Caso de uso, Guardar datos



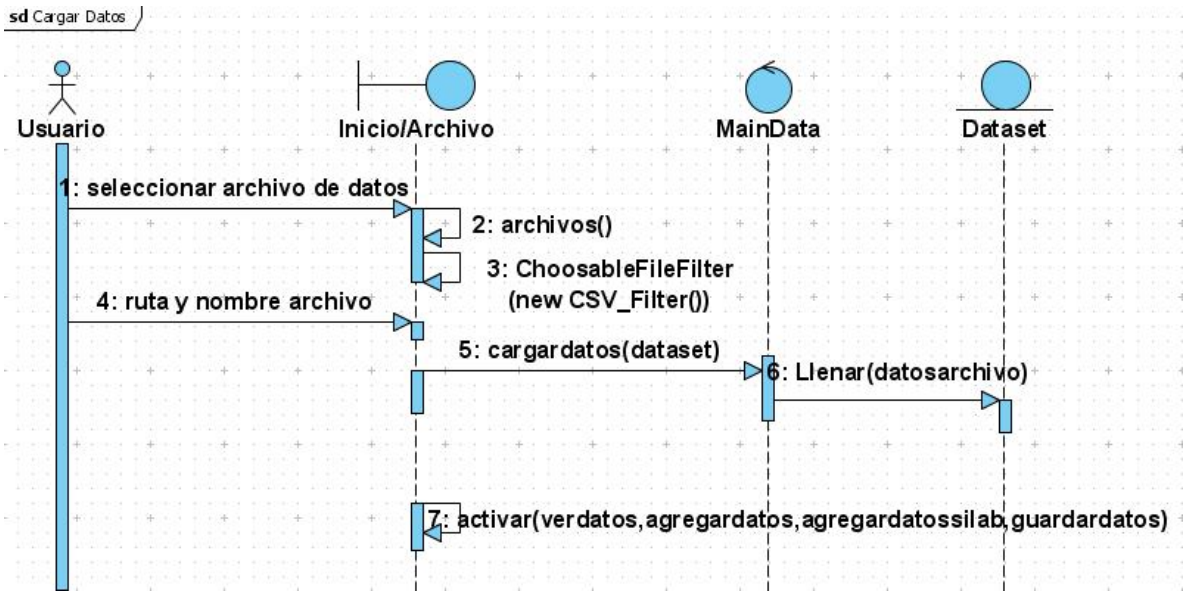
Caso de uso, Guardar resultados



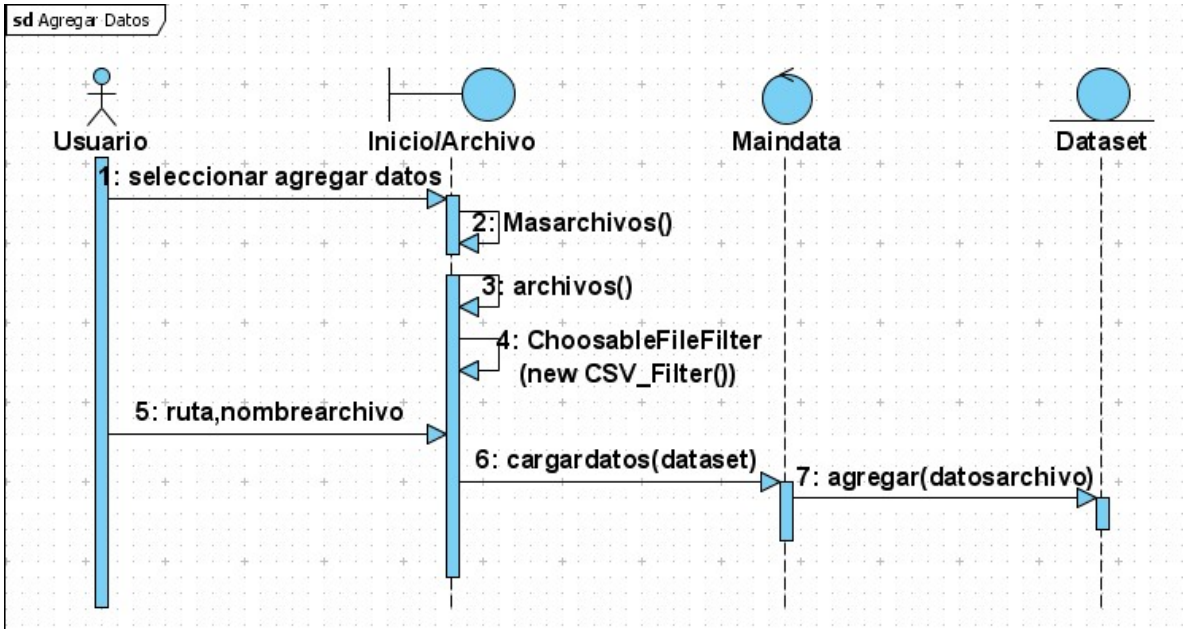
Caso de uso, Ver datos



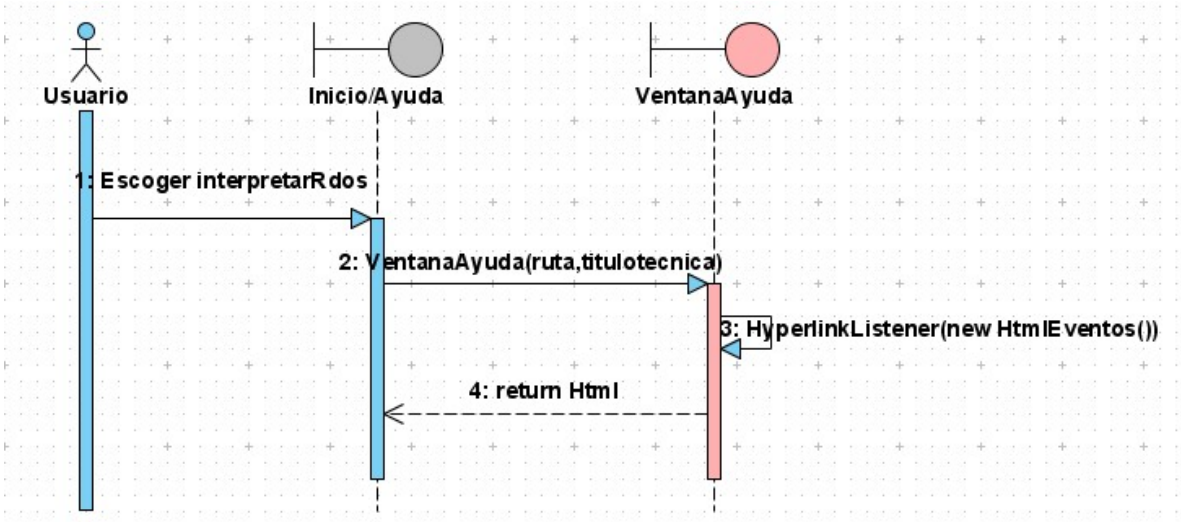
Caso de uso, Cargar Datos



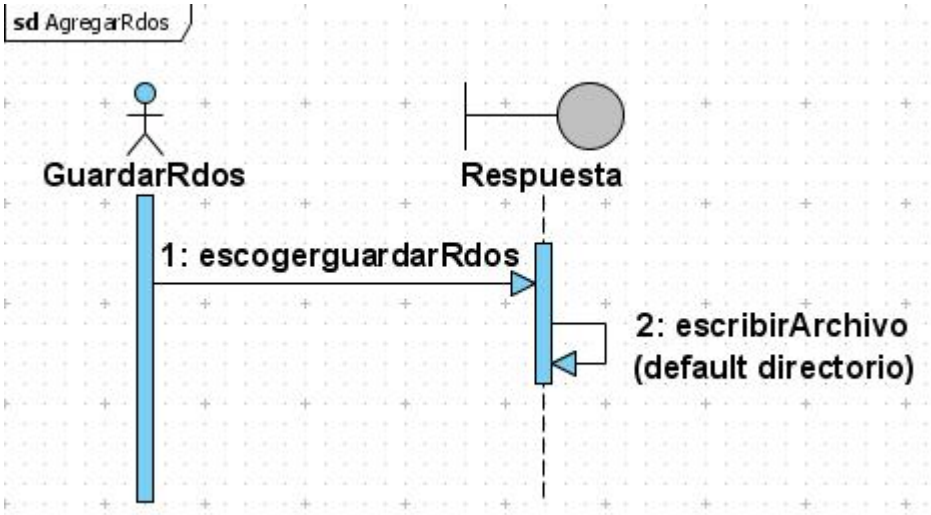
Caso de uso, Agregar Datos



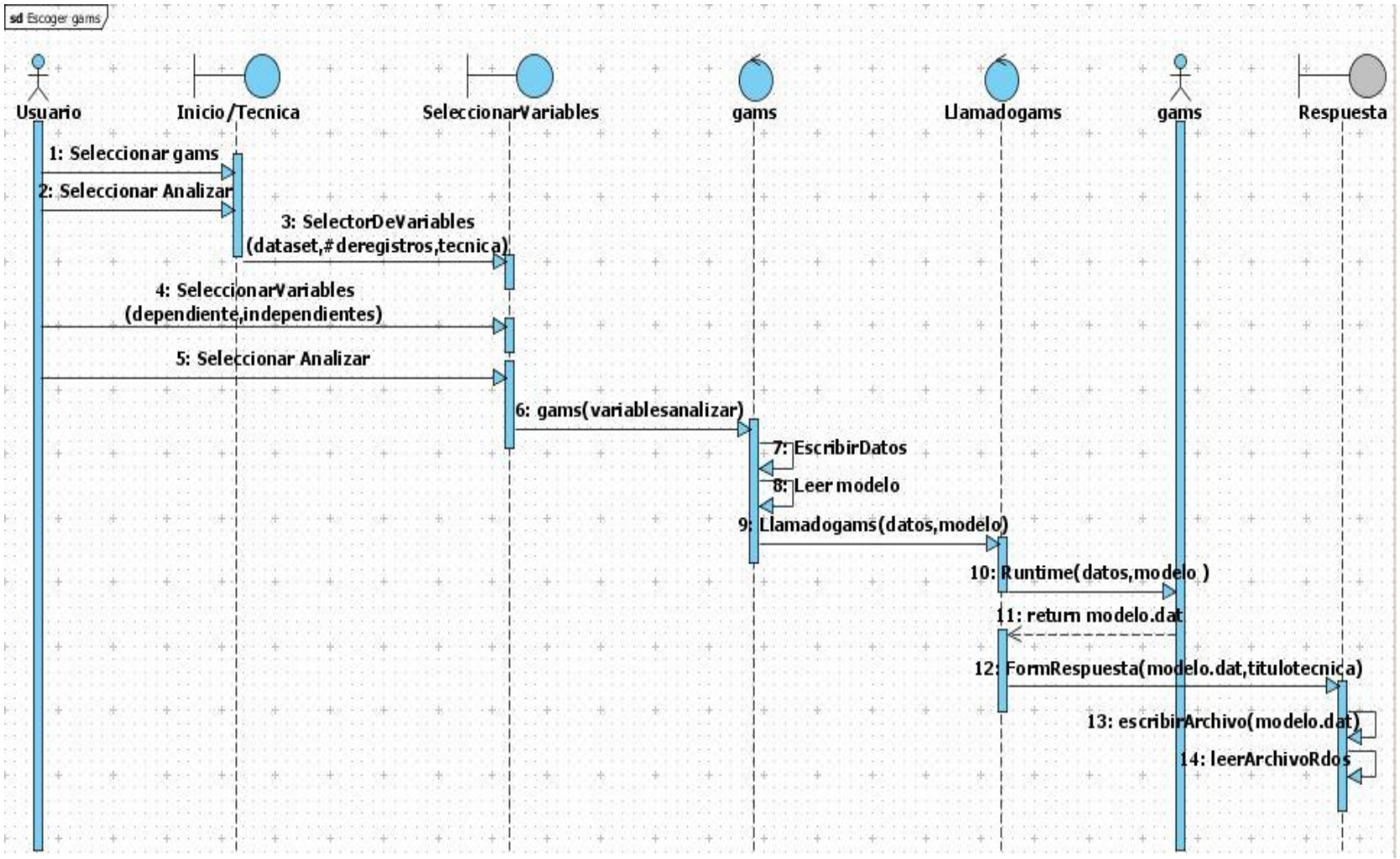
Caso de uso, Interpretar resultados



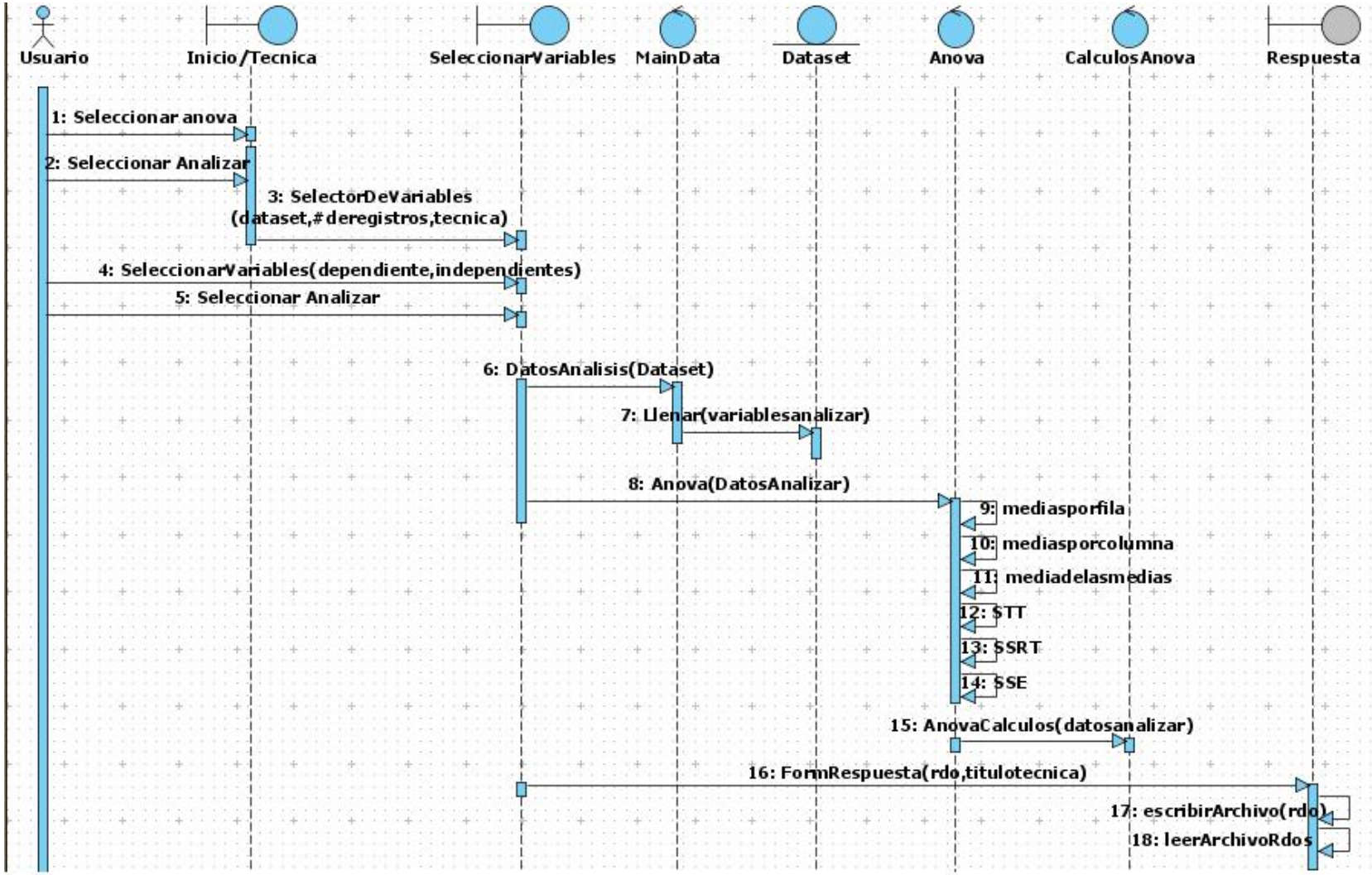
Caso de uso, Agregar resultados



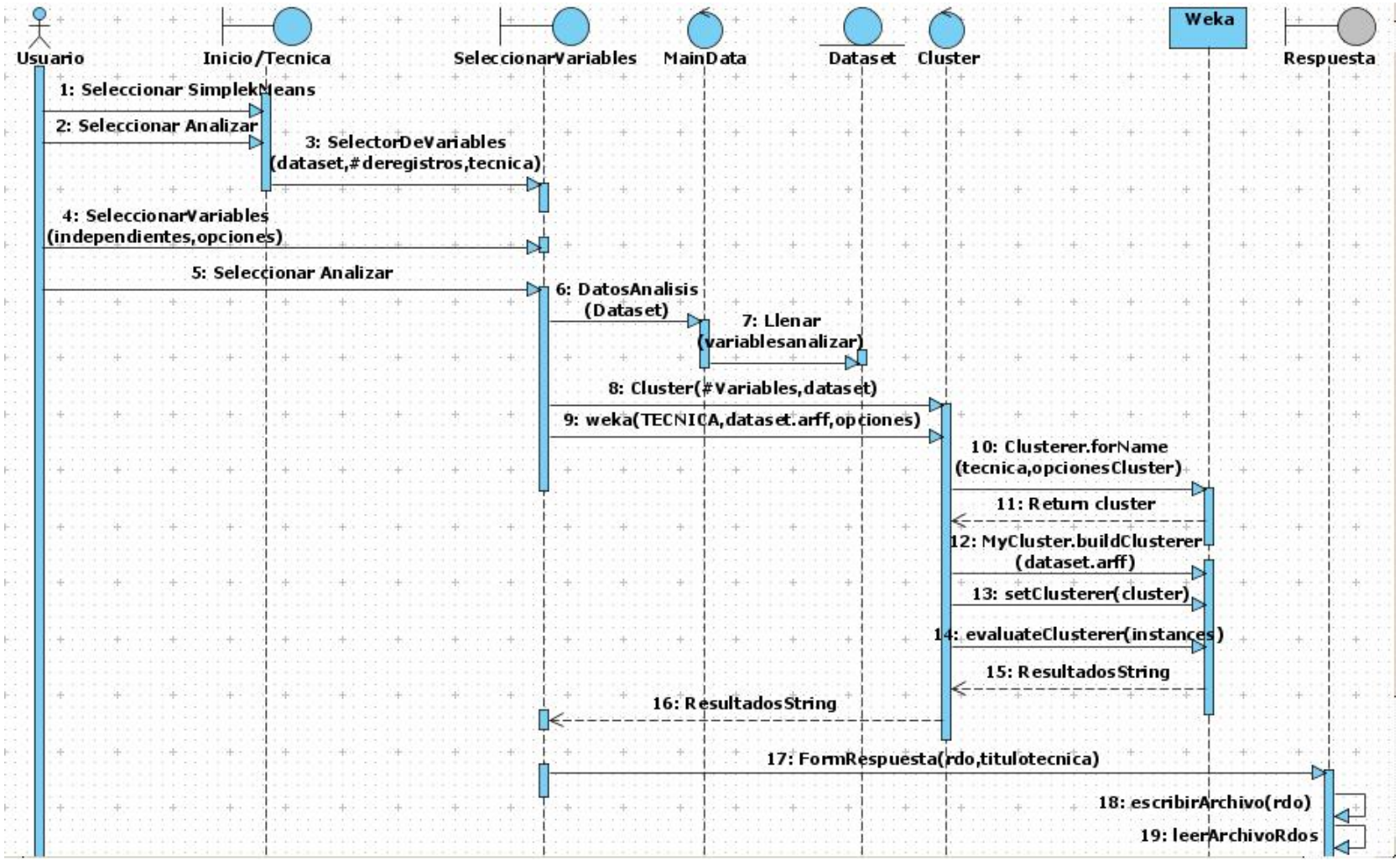
Caso de uso, escoger gams



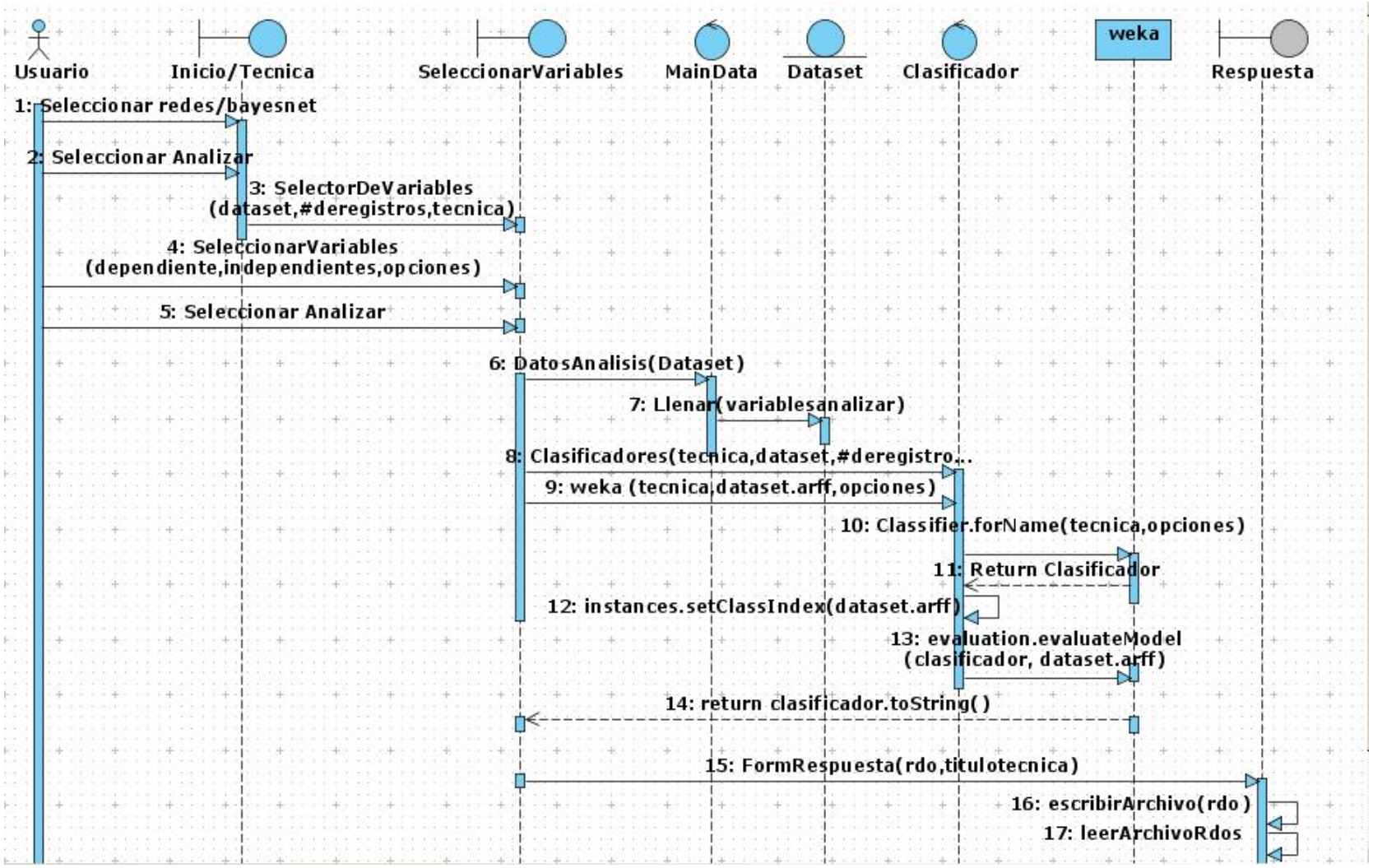
Caso de uso, escoger anova



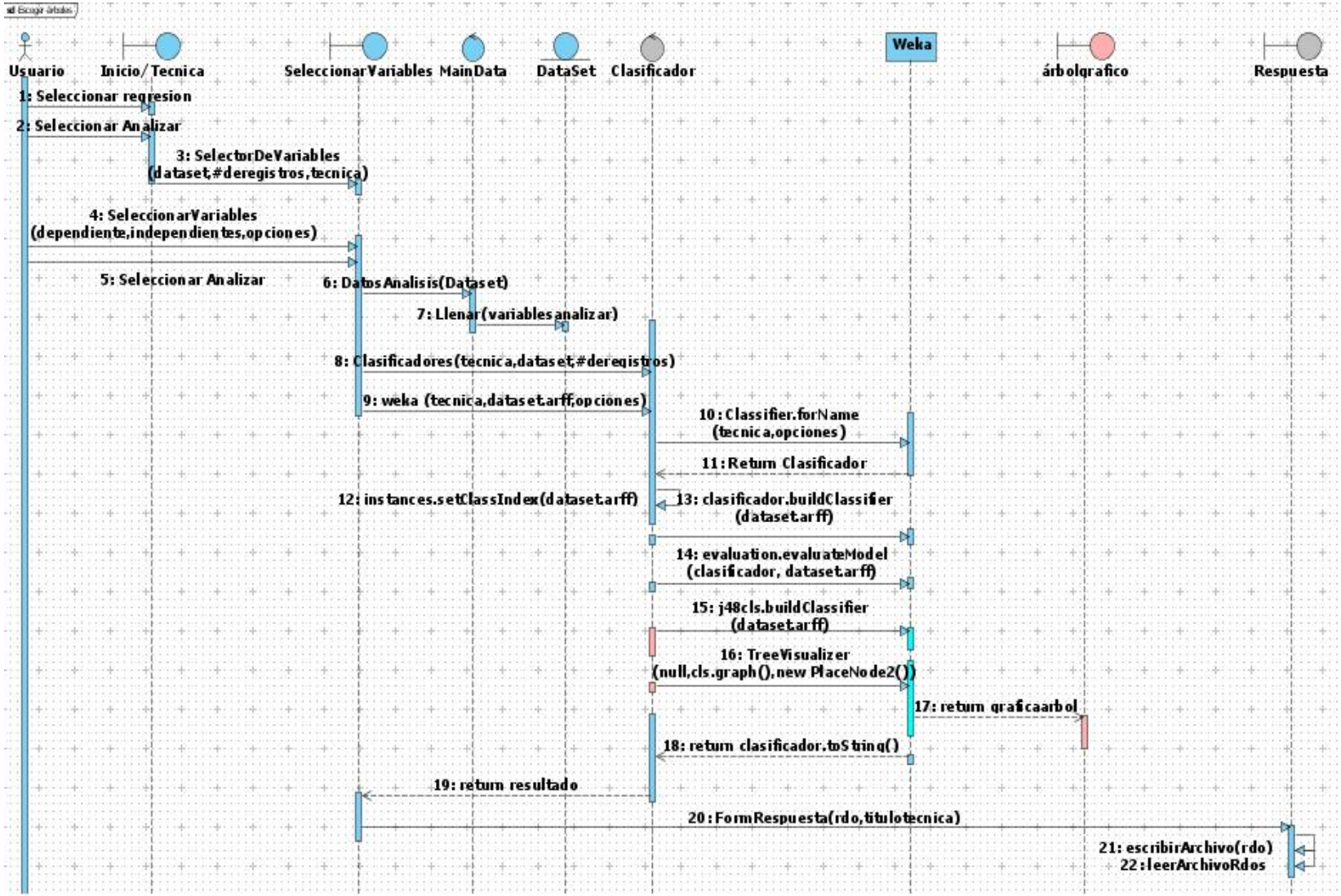
Caso de uso, seleccionar cluster



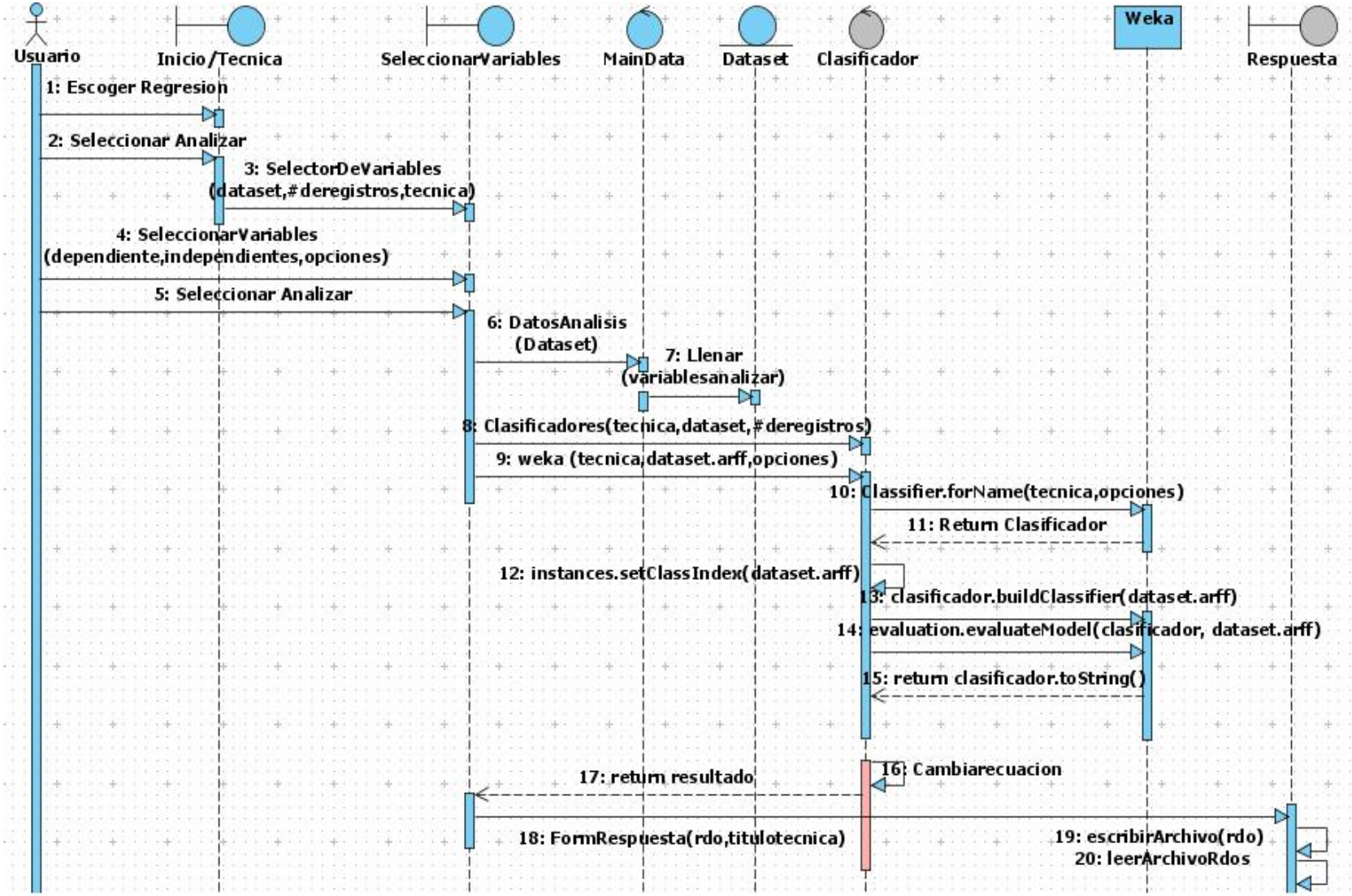
Caso de uso, escoger redes



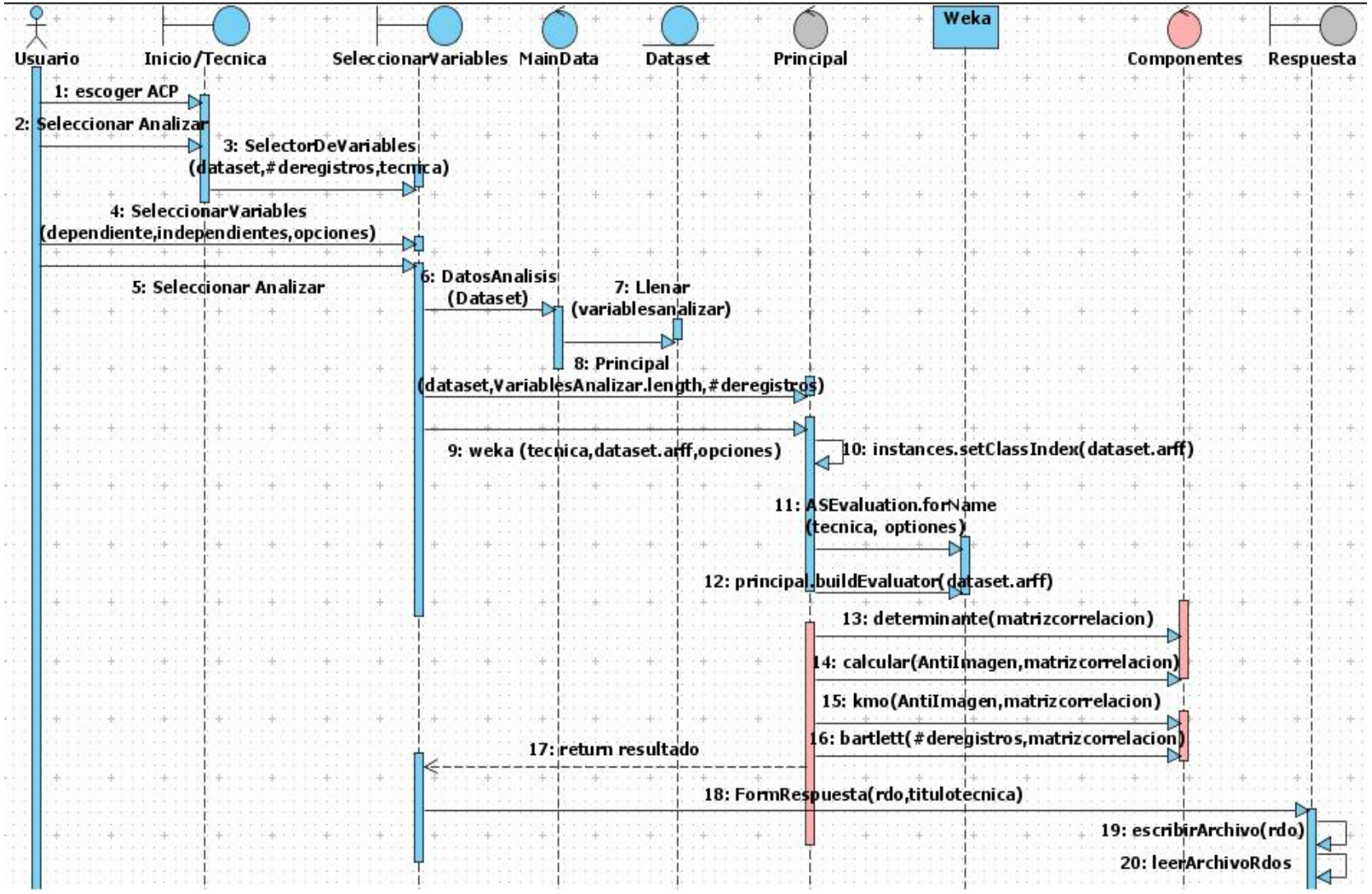
Caso de uso, escoger árboles



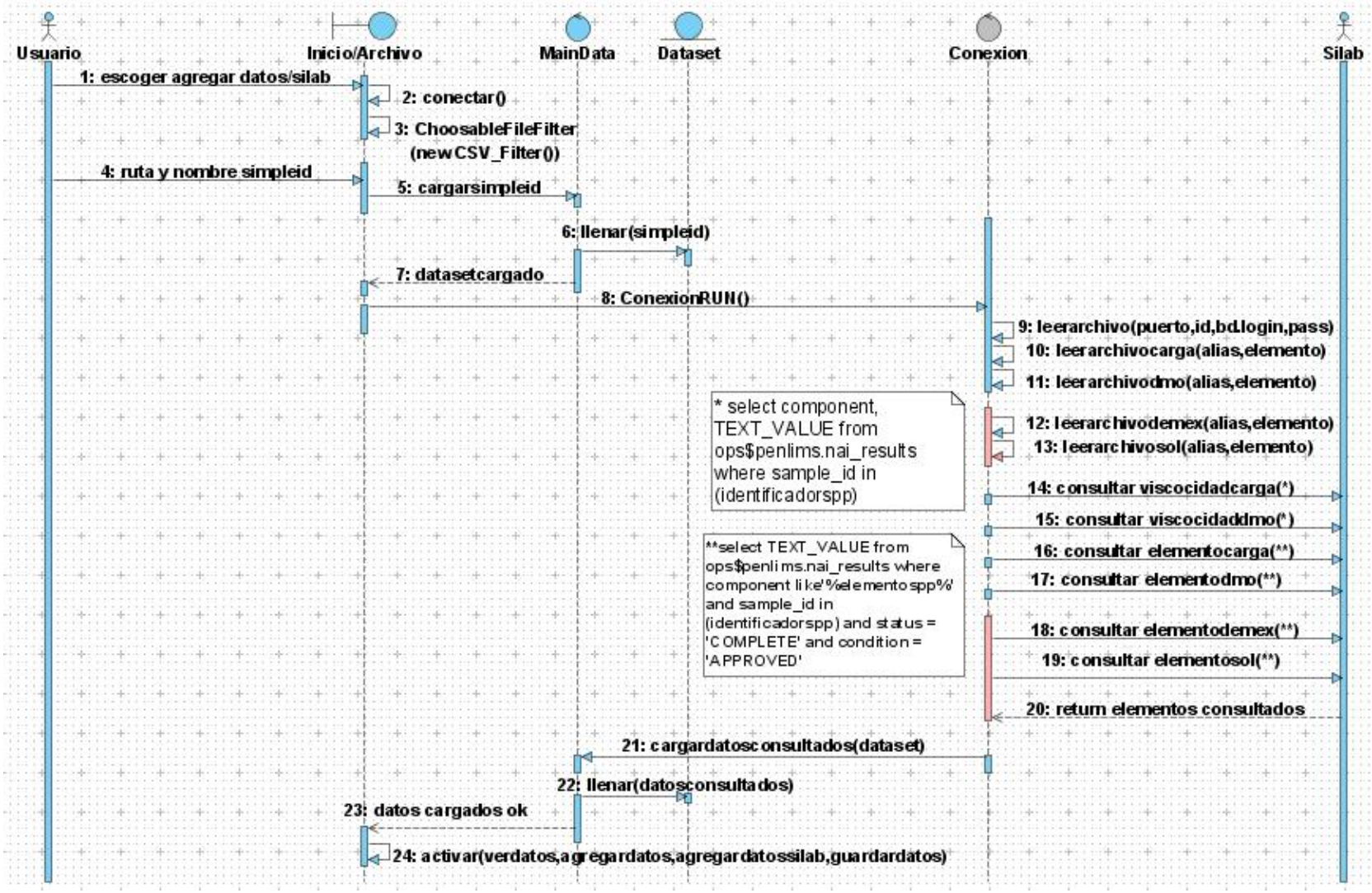
Caso de uso, escoger análisis de regresión



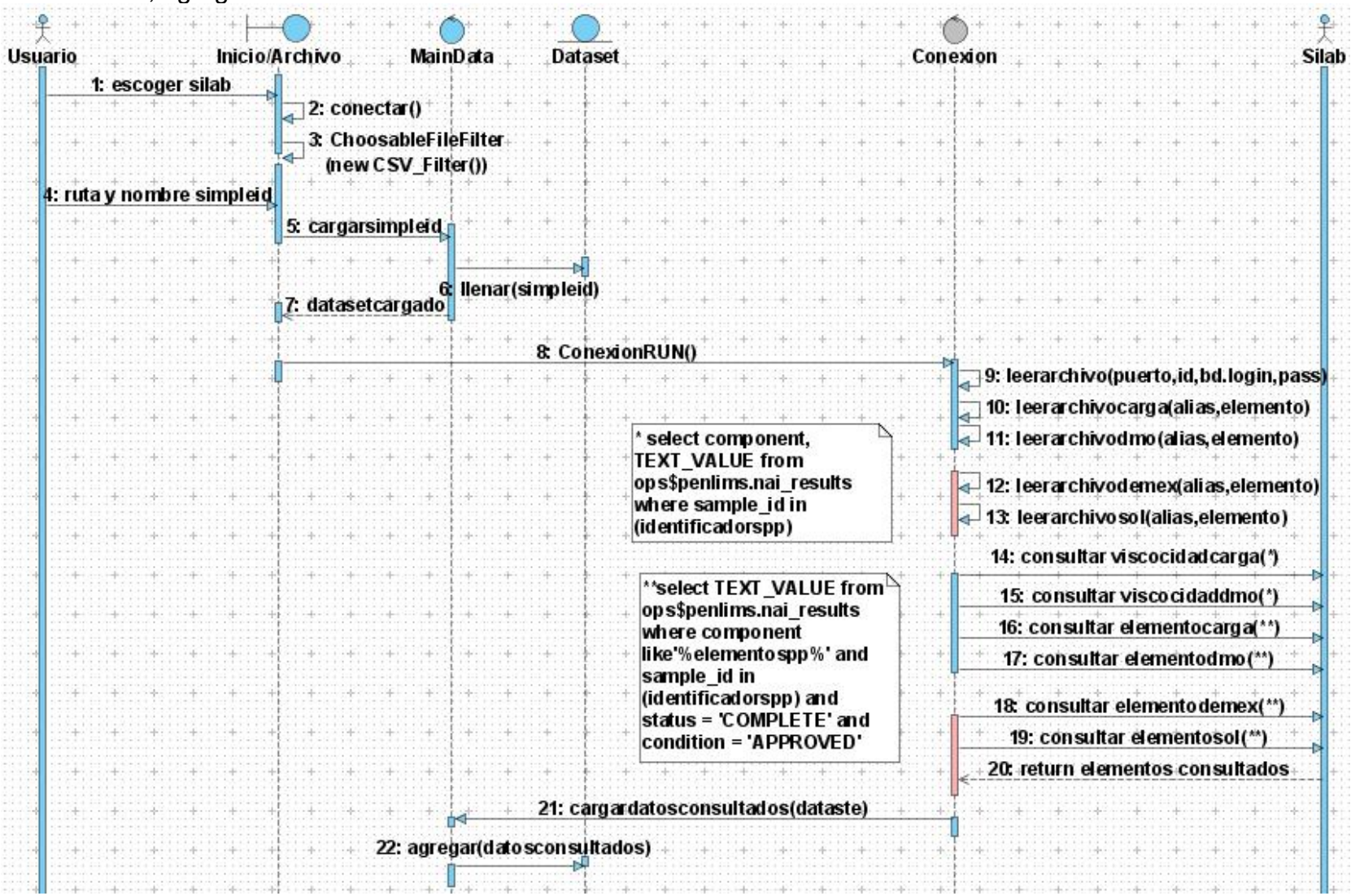
Caso de uso, escoger componentes principales



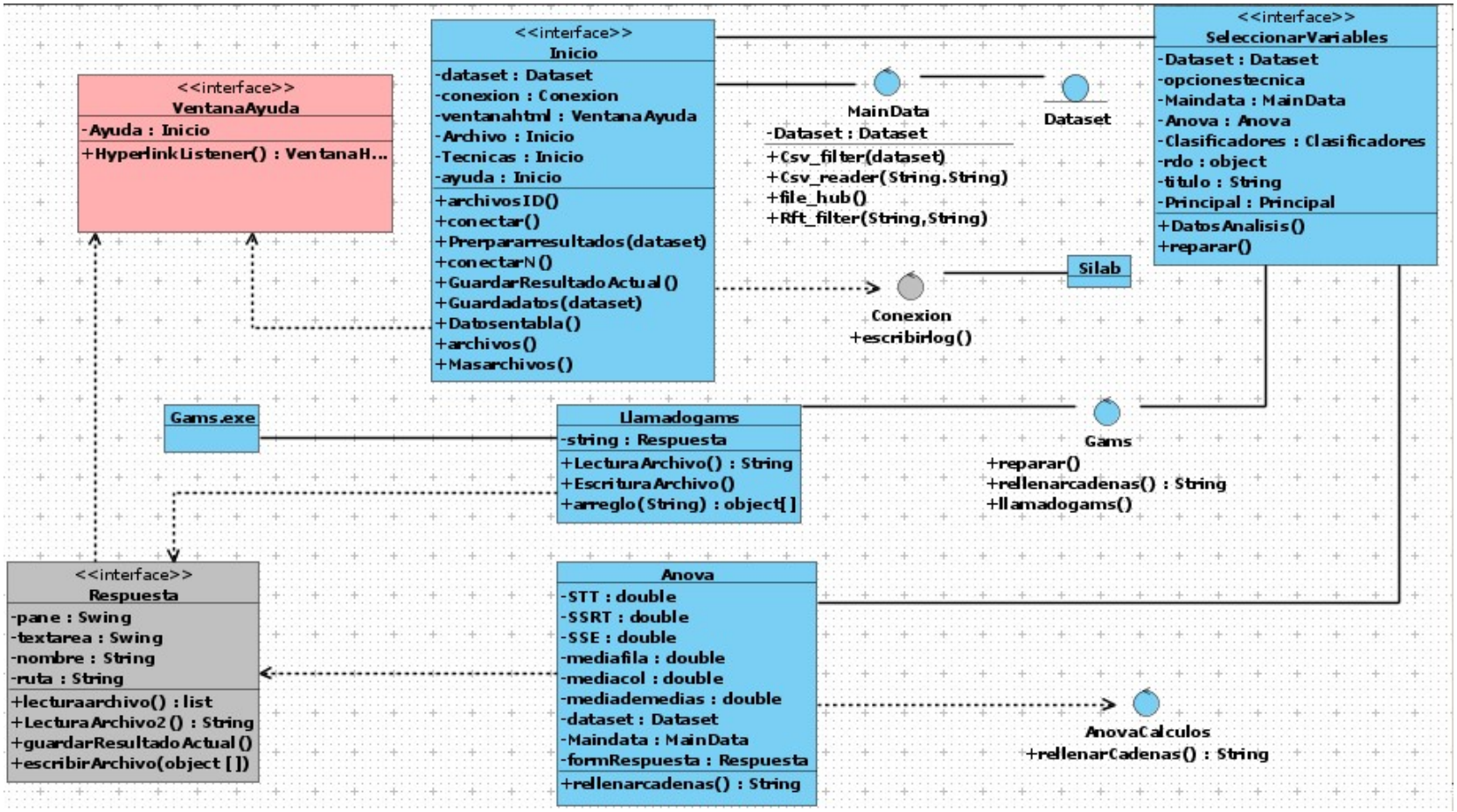
Caso de uso, cargar datos silab

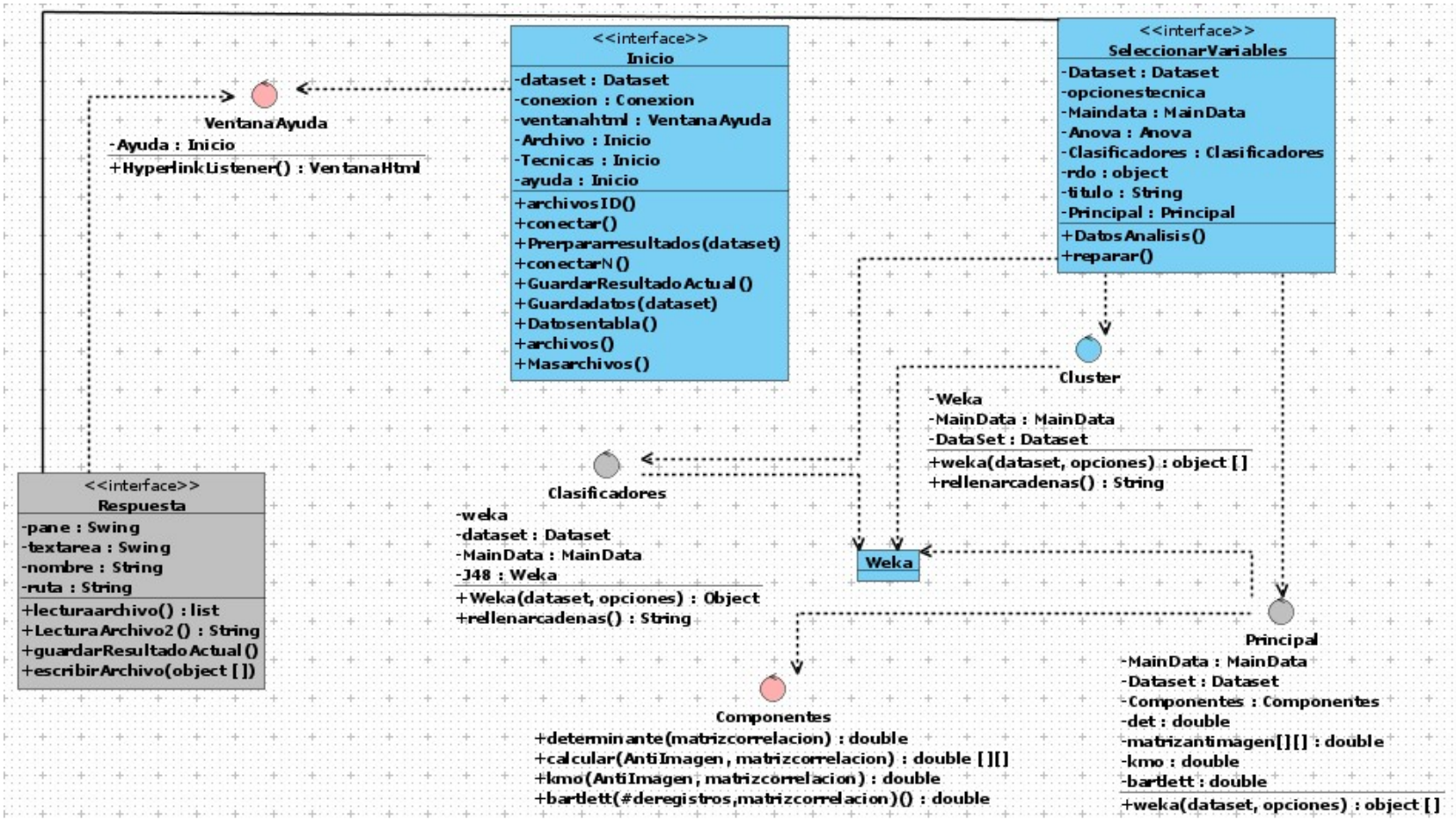


Caso de uso, agregar datos silab



Anexo C. Diagrama de clases




VentanaAyuda
-Ayuda : Inicio
+HyperlinkListener() : VentanaHtml
<<interface>>
Inicio
-dataset : Dataset
-conexion : Conexion
-ventanahtml : VentanaAyuda
-Archivo : Inicio
-Tecnicas : Inicio
-ayuda : Inicio
+archivosID()
+conectar()
+Prerpararresultados(dataset)
+conectarN()
+GuardarResultadoActual()
+Guardadatos(dataset)
+Datosentabla()
+archivos()
+Masarchivos()
<<interface>>
SeleccionarVariables
-Dataset : Dataset
-opcionestecnica
-Maindata : MainData
-Anova : Anova
-Clasificadores : Clasificadores
-rdo : object
-titulo : String
-Principal : Principal
+DatosAnalisis()
+reparar()
Cluster
-Weka
-MainData : MainData
-DataSet : Dataset
+weka(dataset, opciones) : object []
+rellenarcadenas() : String
<<interface>>
Respuesta
-pane : Swing
-textarea : Swing
-nombre : String
-ruta : String
+lecturaarchivo() : list
+LecturaArchivo2() : String
+guardarResultadoActual()
+escribirArchivo(object [])
Clasificadores
-weka
-dataset : Dataset
-MainData : MainData
-J48 : Weka
+Weka(dataset, opciones) : Object
+rellenarcadenas() : String
Weka
Principal
-MainData : MainData
-Dataset : Dataset
-Componentes : Componentes
-det : double
-matrizantimagen[][] : double
-kmo : double
-bartlett : double
+weka(dataset, opciones) : object []
Componentes
+determinante(matrizcorrelacion) : double
+calcular(AntiImagen, matrizcorrelacion) : double [][]
+kmo(AntiImagen, matrizcorrelacion) : double
+bartlett(#deregistros,matrizcorrelacion)() : double

Anexo D. Pruebas adicionales con el prototipo

Las variables C3, i-C4 y n-C4 no pueden ir juntas debido a que los dos valores que toman son constantes a través de toda la sábana de datos, y al aplicar la técnica de componentes principales la correlación entre ellas es 1, imposibilitando el análisis porque sólo puede ser 1 la correlación entre ellas mismas (diagonal principal de la matriz de correlación), es decir, C3 con C3, i-C4 con i-C4 y n-C4 con n-C4.

```
Principal Components Attribute Transformer
-----------------------------------------

Correlation matrix
  1       1      -1
  1       1      -1
 -1      -1       1

Sugerencia:

Se recomienda trabajar con las variables de los componentes cuyos
eigenvalues sean superiores a la unidad.

eigenvalue  proportion  cumulative
  3           1           1         0.577C3+0.577i-C4-0.577n-C4

Eigenvectors
 V1
 0.5774 C3
 0.5774 i-C4
-0.5774 n-C4

 EL DETERMINANTE DE LA MATRIZ ES:
-------------------------------------------------------
0.0
```

```
COEFICIENTE DE ESFERICIDAD DE BARTLETT
------------------------------------------------------
Infinity

KMO
----------------------------------------
NaN


MATRIZ ANTI-IMAGEN DE CORRELACIONES
------------------------------------------------------
C3       i-C4     n-C4
0.0(a)   0.0      0.0
0.0      0.0(a)   0.0
0.0      0.0      0.0(a)
```

C3 i-C4 n-C4 V-carga V50-carga CarAro-carga Penet-carga PIE-carga N2b-carga SatCarga

```
Principal Components Attribute Transformer
-----------------------------------------

Correlation matrix
  1     1    -1    -0.11  -0.03   0.07  -0.09   0.07   0.02  -0.14
  1     1    -1    -0.11  -0.03   0.07  -0.09   0.07   0.02  -0.14
 -1    -1     1     0.11   0.03  -0.07   0.09  -0.07  -0.02   0.14
 -0.11 -0.11  0.11  1      0.44   0.4   -0.55   0.15   0.3   -0.33
 -0.03 -0.03  0.03  0.44   1      0.05  -0.59   0.21   0.26  -0.52
  0.07  0.07 -0.07  0.4    0.05   1     -0.38  -0.11   0.2   -0.41
 -0.09 -0.09  0.09 -0.55  -0.59  -0.38   1     -0.24  -0.55   0.76
  0.07  0.07 -0.07  0.15   0.21  -0.11  -0.24   1      0.25  -0.49
  0.02  0.02 -0.02  0.3    0.26   0.2   -0.55   0.25   1     -0.57
 -0.14 -0.14  0.14 -0.33  -0.52  -0.41   0.76  -0.49  -0.57   1


Sugerencia:
```

Se recomienda trabajar con las variables de los componentes cuyos eigenvalues sean superiores a la unidad.

eigenvalue proportion cumulative

3.4104 0.34104 0.34104 -0.252C3-0.252i-C4+0.252n-C4-0.283V-carga-0.313V50-carga-0.243CarAro-carga+0.45 Penet-carga-0.228PIE-carga-0.33N2b-carga+0.457SatCarga

2.96561 0.29656 0.6376 -0.512C3-0.512i-C4+0.512n-C4+0.262V-carga+0.209V50-carga+0.086CarAro-carga-0.194Penet-carga+0.067PIE-carga+0.168N2b-carga-0.151SatCarga

1.20672 0.12067 0.75827 -0.021C3-0.021i-C4+0.021n-C4-0.288V-carga+0.158V50-carga-0.664CarAro-carga+0.068Penet-carga+0.647PIE-carga+0.094N2b-carga-0.136SatCarga

0.83982 0.08398 0.84226 0.057C3+0.057i-C4-0.057n-C4+0.317V-carga+0.674V50-carga-0.336CarAro-carga-0.097Penet-carga-0.241PIE-carga-0.455N2b-carga+0.219SatCarga

0.66981 0.06698 0.90924 0 C3+0 i-C40 n-C4+0.377V-carga-0.176V50-carga+0.314CarAro-carga+0.165Penet-carga+0.599PIE-carga-0.585N2b-carga-0.006SatCarga

0.53448 0.05345 0.96268 -0.057C3-0.057i-C4+0.057n-C4-0.66V-carga+0.276V50-carga+0.302CarAro-carga-0.088Penet-carga-0.073PIE-carga-0.442N2b-carga-0.423SatCarga

Eigenvectors

V1	V2	V3	V4	V5	V6	
-0.2518	-0.5124	-0.0214	0.0572	0.0002	-0.057	C3
-0.2518	-0.5124	-0.0214	0.0572	0.0002	-0.057	i-C4
0.2518	0.5124	0.0214	-0.0572	-0.0002	0.057	n-C4
-0.2832	0.2618	-0.2883	0.3167	0.3773	-0.6599	V-carga
-0.3126	0.2087	0.1583	0.6739	-0.1756	0.2757	V50-carga
-0.243	0.0858	-0.6636	-0.3363	0.3139	0.302	CarAro-carga
0.4503	-0.1937	0.0684	-0.097	0.1652	-0.0881	Penet-carga
-0.2277	0.0674	0.6466	-0.2409	0.5993	-0.0731	PIE-carga
-0.3305	0.1676	0.0937	-0.4551	-0.5846	-0.4421	N2b-carga
0.4572	-0.1506	-0.136	0.2193	-0.006	-0.423	SatCarga

EL DETERMINANTE DE LA MATRIZ ES:

0.0

COEFICIENTE DE ESFERICIDAD DE BARTLETT

Infinity

KMO

NaN

MATRIZ ANTI-IMAGEN DE CORRELACIONES

C3	i-C4	n-C4	V-carga	V50-carga	CarAro-carga	Penet-carga	PIE-carga	N2b-carga	SatCarga
0.0(a)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0(a)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0(a)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0(a)	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0(a)	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0(a)	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0(a)	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0(a)	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0(a)	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0(a)

Ejercicios con el análisis de componentes principales que cumplen con algunos criterios pero no son completamente buenos. Se uso el conjunto de variables S/C, Trect, D-carga, CCR-carga, Ni-carga, V-carga, V50-carga CarAro-carga, Penet-carga, PIE-carga, N2b-carga, SatCarga, ResCarga, AsfCarga y C3

Principal Components Attribute Transformer

Correlation matrix

1	0.07	0.03	0	0.02	-0.01	0	0.02	-0.01	-0.01	0.02	-0.02	0	0	0.01	
0.07	1	-0.09	0.07	0.1	0.1	-0.11	-0.03	0.07	-0.09	0.07	0.02	-0.14	0.01	-0.05	
0.03	-0.09	1	0.06	0.02	0.05	0.07	-0.01	0.09	0	-0.04	0.03	0.01	-0.05	0.05	
0	0.07	0.06	1	0.88	0.68	0.63	0.55	0.82	-0.71	0.21	0.38	-0.72	-0.18	0.77	
0.02	0.1	0.02	**0.88**	1	0.59	0.52	0.57	0.59	-0.79	0.34	0.37	-0.75	-0.06	0.81	
-0.01	0.1	0.05	**0.68**	**0.59**	1	0.6	0.43	0.56	-0.67	-0.15	0.37	-0.52	-0.4	0.68	
0	-0.11	0.07	**0.63**	**0.52**	**0.6**	1	0.44	0.4	-0.55	0.15	0.3	-0.33	-0.17	0.65	
0.02	-0.03	-0.01	**0.55**	**0.57**	**0.43**	**0.44**	1	0.05	-0.59	0.21	0.26	-0.52	0.02	0.44	
-0.01	0.07	0.09	**0.82**	**0.59**	**0.56**	**0.4**	0.05	1	-0.38	-0.11	0.2	-0.41	-0.37	0.61	
-0.01	-0.09	0	**-0.71**	**-0.79**	**-0.67**	**-0.55**	**-0.59**	**-0.38**	1	-0.24	-0.55	0.76	-0.08	-0.61	

0.02	0.07	-0.04	0.21	**0.34**	-0.15	0.15	0.21	-0.11	-0.24	1	0.25	-0.49	0.54	0.01
-0.02	0.02	0.03	**0.38**	**0.37**	**0.37**	**0.3**	0.26	0.2	**-0.55**	0.25	1	-0.57	0.23	0.23
0	-0.14	0.01	**-0.72**	**-0.75**	**-0.52**	**-0.33**	**-0.52**	**-0.41**	**0.76**	**-0.49**	**-0.57**	1	-0.28	-0.4
0	0.01	-0.05	-0.18	-0.06	**-0.4**	-0.17	0.02	**-0.37**	-0.08	**0.54**	0.23	-0.28	1	-0.45
0.01	-0.05	0.05	**0.77**	**0.81**	**0.68**	**0.65**	**0.44**	**0.61**	**-0.61**	0.01	0.23	**-0.4**	**-0.45**	1

Sugerencia:

Se recomienda trabajar con las variables de los componentes cuyos eigenvalues sean superiores a la unidad.

eigenvalue proportion cumulative

6.06414 0.40428 0.40428 0.003S/C+0.028C3+0.021Trect+0.382D-carga+0.372CCR-carga+0.32 Ni-carga+0.286V-carga+0.255V50-carga+0.27 CarAro-carga-0.35Penet-carga+0.097PIE-carga+0.211N2b-carga-0.318SatCarga-0.058ResCarga+0.334AsfCarga

2.29857 0.15324 0.55751 0.01 S/C+0.07 C3-0.081Trect-0.057D-carga+0.055CCR-carga-0.214Ni-carga-0.096V-carga+0.136V50-carga-0.274CarAro-carga-0.143Penet-carga+0.506PIE-carga+0.254N2b-carga-0.307SatCarga+0.584ResCarga-0.24AsfCarga

1.16108 0.07741 0.63492 -0.237S/C-0.795C3+0.349Trect-0.059D-carga-0.061CCR-carga-0.056Ni-carga+0.27 V-carga+0.204V50-carga-0.199CarAro-carga-0.002Penet-carga+0.033PIE-carga+0.055N2b-carga+0.105SatCarga+0.063ResCarga+0.074AsfCarga

1.03427 0.06895 0.70387 -0.718S/C-0.036C3-0.67Trect-0.03D-carga-0.014CCR-carga+0.062Ni-carga+0.003V-carga+0.106V50-carga-0.086CarAro-carga-0.042Penet-carga-0.077PIE-carga-0.015N2b-carga+0.008SatCarga-0.046ResCarga+0.015AsfCarga

0.98934 0.06596 0.76983 0.571S/C-0.121C3-0.475Trect-0.074D-carga+0.029CCR-carga-0.006Ni-carga+0.17 V-carga+0.417V50-carga-0.339CarAro-carga-0.042Penet-carga+0.016PIE-carga-0.257N2b-carga+0.131SatCarga-0.094ResCarga+0.124AsfCarga

0.84309 0.05621 0.82603 -0.055S/C-0.163C3-0.106Trect+0.229D-carga+0.247CCR-carga-0.36Ni-carga+0.012V-carga-0.193V50-carga+0.314CarAro-carga+0.234Penet-carga+0.464PIE-carga-0.528N2b-carga+0.013SatCarga+0.03 ResCarga+0.174AsfCarga

0.7792 0.05195 0.87798 -0.309S/C+0.441C3+0.428Trect-0.029D-carga+0.092CCR-carga+0.002Ni-carga-0.007V-carga+0.461V50-carga-0.302CarAro-carga-0.023Penet-carga+0.077PIE-carga-0.444N2b-carga+0.049SatCarga-0.089ResCarga+0.032AsfCarga

0.5671 0.03781 0.91579 0.051S/C-0.298C3+0.013Trect+0.119D-carga+0.166CCR-carga-0.061Ni-carga-0.734V-carga+0.298V50-carga+0.121CarAro-carga-0.066Penet-carga-0.326PIE-carga-0.149N2b-carga-0.285SatCarga+0.023ResCarga-0.084AsfCarga

0.37844 0.02523 0.94102 0.02 S/C-0.002C3-0.024Trect+0.142D-carga-0.189CCR-carga+0.153Ni-carga+0.359V-carga+0.029V50-carga+0.253CarAro-carga-0.121Penet-carga-0.345PIE-carga-0.427N2b-carga-0.07SatCarga+0.524ResCarga-0.362AsfCarga

0.32018 0.02135 0.96236 0.003S/C+0.086C3-0.032Trect+0.31 D-carga-0.235CCR-carga-0.238Ni-carga+0.13 V-carga+0.499V50-carga+0.276CarAro-carga+0.503Penet-carga+0.029PIE-carga+0.303N2b-carga+0.016SatCarga-0.151ResCarga-0.269AsfCarga

Eigenvectors

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	
0.0027	0.0101	-0.2375	**-0.7177**	0.5712	-0.0549	-0.3086	0.0514	0.0205	0.003	S/C
0.0279	0.0697	**-0.7953**	-0.0364	-0.1212	-0.1626	0.441	-0.2976	-0.0024	0.0862	C3
0.0208	-0.0813	0.3489	**-0.6705**	-0.4751	-0.1061	0.4276	0.0134	-0.0241	-0.0318	Trect
0.3818	-0.0573	-0.0592	-0.0304	-0.0738	0.2295	-0.0287	0.1187	0.1417	0.3097	D-carga

0.3717	0.0553	-0.0612	-0.0137	0.0293	0.2472	0.0923	0.166	-0.1887	-0.2348	CCR-carga
0.3204	-0.2137	-0.0555	0.062	-0.0059	-0.3597	0.0017	-0.0608	0.1527	-0.2381	Ni-carga
0.2857	-0.0963	0.2705	0.003	0.1703	0.0124	-0.0071	-0.7338	0.3586	0.1303	V-carga
0.2548	0.1357	0.204	0.1063	0.4174	-0.193	0.4605	0.2984	0.0292	0.4994	V50-carga
0.27	**-0.2735**	-0.1995	-0.0857	-0.3385	0.3137	-0.3018	0.1212	0.2527	0.2757	CarAro-carga
-0.3499	-0.1428	-0.0022	-0.0421	-0.0421	0.2345	-0.0234	-0.0656	-0.1212	0.5034	Penet-carga
0.0968	**0.5058**	0.0335	-0.0774	0.0161	0.4638	0.0773	-0.3257	-0.3448	0.0295	PIE-carga
0.2113	**0.2539**	0.0551	-0.0152	-0.2566	-0.528	-0.4438	-0.1486	-0.4272	0.3032	N2b-carga
-0.3182	-0.3073	0.1048	0.0085	0.1307	0.013	0.0495	-0.2847	-0.0699	0.016	SatCarga
-0.0581	**0.5842**	0.063	-0.0461	-0.0944	0.0304	-0.0889	0.023	0.5239	-0.1508	ResCarga
0.3342	-0.2396	0.0737	0.0147	0.1236	0.1743	0.0317	-0.0839	-0.3624	-0.2688	AsfCarga

EL DETERMINANTE DE LA MATRIZ ES:

4.757733364886142E-7

COEFICIENTE DE ESFERICIDAD DE BARTLETT

2897.106531714821

KMO

0.6617667999396619

MATRIZ ANTI-IMAGEN DE CORRELACIONES

S/C	C3	Trect	D-carga	CCR-carga	Ni-carga	V-carga	V50-carga	CarAro-carga	Penet-carga	PIE-carga	N2b-carga	SatCarga	ResCarga	AsfCarga
0.184(a)	-0.085	-0.04	0.076	-0.042	0.031	-0.05	-0.076	-0.072	0.014	-0.047	0.012	0.019	0.0070	-0.0080
-0.085	**0.257(a)**	0.089	-0.151	-0.094	-0.209	0.175	0.181	0.15	0.014	0.015	0.034	-0.126	0.085	0.246
-0.04	0.089	**0.684(a)**	-0.013	-0.014	-0.029	-0.015	0.012	0.0	-0.033	0.01	-0.041	-0.037	-0.0050	0.031
0.076	-0.151	-0.013	**0.602(a)**	-0.502	0.148	-0.636	-0.921	-0.972	0.031	-0.531	-0.056	0.383	0.04	-0.019
-0.042	-0.094	-0.014	-0.502	**0.745(a)**	0.113	0.507	0.356	0.406	0.325	0.052	0.211	0.058	-0.162	-0.643
0.031	-0.209	-0.029	0.148	0.113	**0.809(a)**	-0.345	-0.141	-0.151	0.273	0.2	-0.028	0.476	0.358	-0.193
-0.05	0.175	-0.015	-0.636	0.507	-0.345	**0.578(a)**	0.494	0.561	0.158	0.056	0.0070	-0.465	-0.201	-0.255
-0.076	0.181	0.012	-0.921	0.356	-0.141	0.494	**0.428(a)**	0.945	-0.035	0.564	0.085	-0.263	0.024	0.066
-0.072	0.15	0.0	-0.972	0.406	-0.151	0.561	0.945	**0.457(a)**	-0.099	0.594	0.05	-0.301	0.013	0.04
0.014	0.014	-0.033	0.031	0.325	0.273	0.158	-0.035	-0.099	**0.902(a)**	-0.249	0.24	-0.159	0.177	-0.072
-0.047	0.015	0.01	-0.531	0.052	0.2	0.056	0.564	0.594	-0.249	**0.455(a)**	0.014	0.127	-0.101	0.079
0.012	0.034	-0.041	-0.056	0.211	-0.028	0.0070	0.085	0.05	0.24	0.014	**0.876(a)**	0.224	-0.119	-0.121
0.019	-0.126	-0.037	0.383	0.058	0.476	-0.465	-0.263	-0.301	-0.159	0.127	0.224	**0.766(a)**	0.235	-0.288

0.0070	0.085	-0.0050	0.04-	0.162	0.358	-0.201	0.024	0.013	0.177	-0.101	-0.119	0.235	**0.677(a)** 0.396
-0.008	0.246	0.031	-0.019	-0.643	-0.193	-0.255	0.066	0.04	-0.072	0.079	-0.121 -	0.288	0.396 **0.803(a)**

Eliminando la variable C3 se tiene

Principal Components Attribute Transformer

Correlation matrix

1 0.03 0 0.02 -0.01 0 0.02 -0.01 -0.01 0.02 -0.02 0 0 0.01

0.03 1 0.06 0.02 0.05 0.07 -0.01 0.09 0 -0.04 0.03 0.01 -0.05 0.05

0 0.06 1 0.88 0.68 0.63 0.55 0.82 -0.71 0.21 0.38 -0.72 -0.18 0.77

0.02 0.02 **0.88** 1 0.59 0.52 0.57 0.59 -0.79 0.34 0.37 -0.75 -0.06 0.81

-0.01 0.05 **0.68** **0.59** 1 0.6 0.43 0.56 -0.67 -0.15 0.37 -0.52 -0.4 0.68

0 0.07 **0.63** **0.52** **0.6** 1 0.44 0.4 -0.55 0.15 0.3 -0.33 -0.17 0.65

0.02 -0.01 **0.55** **0.57** **0.43** **0.44** 1 0.05 -0.59 0.21 0.26 -0.52 0.02 0.44

-0.01 0.09 **0.82** **0.59** **0.56** **0.4** 0.05 1 -0.38 -0.11 0.2 -0.41 -0.37 0.61

-0.01 0 **-0.71** **-0.79** **-0.67** **-0.55** **-0.59** **-0.38** 1 -0.24 -0.55 0.76 -0.08 -0.61

0.02 -0.04 0.21 **0.34** -0.15 0.15 0.21 -0.11 -0.24 1 0.25 -0.49 0.54 0.01

-0.02 0.03 **0.38** **0.37** **0.37** **0.3** 0.26 0.2 **-0.55** 0.25 1 -0.57 0.23 0.23

0 0.01 **-0.72** **-0.75** **-0.52** **-0.33** **-0.52** **-0.41** **0.76** **-0.49** **-0.57** 1 -0.28 -0.4

0 -0.05 -0.18 -0.06 **-0.4** -0.17 0.02 -0.37 -0.08 **0.54** 0.23 -0.28 1 -0.45

0.01 0.05 **0.77** **0.81** **0.68** **0.65** **0.44** **0.61** **-0.61** 0.01 0.23 **-0.4** **-0.45** 1

Sugerencia:

Se recomienda trabajar con las variables de los componentes cuyos eigenvalues sean superiores a la unidad.

eigenvalue proportion cumulative

6.06021 0.43287 0.43287 0.002S/C+0.021Trect+0.382D-carga+0.372CCR-carga+0.32 Ni-carga+0.287V-carga+0.255V50-carga+0.27 CarAro-carga-0.35Penet-carga+0.096PIE-carga+0.211N2b-carga-0.318SatCarga-0.058ResCarga+0.335AsfCarga

2.29239 0.16374 0.59661 0.007S/C-0.077Trect-0.058D-carga+0.055CCR-carga-0.216Ni-carga-0.089V-carga+0.14 V50-carga-0.277CarAro-carga-0.143Penet-carga+0.507PIE-carga+0.256N2b-carga-0.306SatCarga+0.587ResCarga-0.236AsfCarga

1.03464 0.0739 0.67052 -0.679S/C-0.713Trect-0.026D-carga-0.007CCR-carga+0.063Ni-carga-0.015V-carga+0.096V50-carga-0.074CarAro-carga-0.042Penet-carga-0.077PIE-carga-0.024N2b-carga+0.002SatCarga-0.051ResCarga+0.013AsfCarga

0.99518 0.07108 0.7416 0.665S/C-0.554Trect-0.041D-carga+0.052CCR-carga-0.009Ni-carga+0.08 V-carga+0.307V50-carga-0.234CarAro-carga-0.03Penet-carga+0.022PIE-carga-0.252N2b-carga+0.086SatCarga-0.094ResCarga+0.101AsfCarga

0.89422 0.06387 0.80547 -0.204S/C+0.369Trect-0.196D-carga-0.123CCR-carga+0.144Ni-carga+0.276V-carga+0.558V50-carga-0.54CarAro-carga-0.128Penet-carga-0.152PIE-carga+0.004N2b-carga+0.16 SatCarga-0.059ResCarga+0.022AsfCarga

0.83081 0.05934 0.86482 -0.231S/C+0.162Trect+0.156D-carga+0.234CCR-carga-0.313Ni-carga+0.11 V-carga+0.123V50-carga+0.047CarAro-carga+0.188Penet-carga+0.444PIE-carga-0.662N2b-carga+0.085SatCarga-0.016ResCarga+0.192AsfCarga

0.60557 0.04326 0.90807 0.014S/C-0.117Trect-0.085D-carga-0.172CCR-carga-0.025Ni-carga+0.747V-carga-0.36V50-carga-0.03CarAro-carga+0.101Penet-carga+0.296PIE-carga+0.251N2b-carga+0.291SatCarga+0.029ResCarga+0.121AsfCarga

0.37844 0.02703 0.9351 -0.02S/C+0.024Trect-0.141D-carga+0.189CCR-carga-0.155Ni-carga-0.359V-carga-0.027V50-carga-0.251CarAro-carga+0.122Penet-carga+0.343PIE-carga+0.429N2b-carga+0.07 SatCarga-0.523ResCarga+0.363AsfCarga

0.32825 0.02345 0.95855 0.009S/C-0.046Trect+0.309D-carga-0.115CCR-carga-0.45Ni-carga+0.088V-carga+0.492V50-carga+0.295CarAro-carga+0.412Penet-carga-

0.157PIE-carga+0.357N2b-carga+0.13 SatCarga+0.01 ResCarga-0.103AsfCarga

Eigenvectors

V1	V2	V3	V4	V5	V6	V7	V8	V9	
0.0023	0.0065	**-0.6794**	0.6653	-0.204	-0.2307	0.0138	-0.0201	0.0088	S/C
0.0214	-0.0769	**-0.7126**	-0.5539	0.3694	0.1618	-0.1168	0.0236	-0.0456	Trect
0.3819	-0.0575	-0.0256	-0.0406	-0.1962	0.1565	-0.0849	-0.1405	0.3093	D-carga
0.3716	0.0547	-0.0072	0.0519	-0.1228	0.2336	-0.1717	0.189	-0.1155	CCR-carga
0.3204	-0.2159	0.0631	-0.009	0.1444	-0.3129	-0.0252	-0.1555	-0.4502	Ni-carga
0.2866	-0.0895	-0.0147	0.0802	0.276	0.1097	0.7465	-0.3595	0.0876	V-carga
0.2552	0.1403	0.0957	0.3066	0.5578	0.123	-0.3602	-0.0274	0.4923	V50-carga
0.2699	**-0.2766**	-0.0735	-0.2335	-0.5396	0.047	-0.0296	-0.2512	0.2948	CarAro-carga
-0.3498	-0.1431	-0.0416	-0.0296	-0.1282	0.1883	0.1008	0.1222	0.4119	Penet-carga
0.0964	**0.5072**	-0.0771	0.0223	-0.152	0.4438	0.2957	0.3428	-0.1568	PIE-carga
0.2114	**0.2562**	-0.0241	-0.2518	0.0044	-0.6616	0.2511	0.4286	0.3566	N2b-carga
-0.3178	-0.3059	0.002	0.0862	0.1601	0.0846	0.2911	0.0704	0.13	SatCarga
-0.0583	**0.5866**	-0.051	-0.0943	-0.059	-0.0159	0.0289	-0.5229	0.0104	ResCarga
0.3349	-0.2365	0.0128	0.1009	0.0221	0.1915	0.1211	0.363	-0.1035	AsfCarga

EL DETERMINANTE DE LA MATRIZ ES:

5.786739808038289E-7

COEFICIENTE DE ESFERICIDAD DE BARTLETT

2858.1427915763943

KMO

0.6676106787669152

MATRIZ ANTI-IMAGEN DE CORRELACIONES

S/C	Trect	D-carga	CCR-carga	Ni-carga	V-carga	V50-carga	CarAro-carga	Penet-carga	PIE-carga	N2b-carga	SatCarga	ResCarga	AsfCarga
0.125(a)	-0.033	0.064	-0.05	0.014	-0.036	-0.062	-0.06	0.015	-0.046	0.015	0.0080	0.014	0.012
-0.033	**0.805(a)**	0.0-	0.0060	-0.011	-0.032	-0.0030	-0.013	-0.034	0.0090	-0.044	-0.026	-0.012	0.01
0.064	0.0	**0.604(a)**	-0.525	0.12	-0.627	-0.919	-0.971	0.033	-0.535	-0.052	0.371	0.054	0.018
-0.05	-0.0060	-0.525	**0.735(a)**	0.096	0.534	0.381	0.427	0.328	0.054	0.215	0.047	-0.155	-0.643
0.014	-0.011	0.12	0.096	**0.826(a)**	-0.32	-0.107	-0.124	0.282	0.207	-0.021	0.463	0.386	-0.149
-0.036	-0.032	-0.627	0.534	-0.32	**0.581(a)**	0.477	0.549	0.158	0.055	0.0010	-0.454	-0.221	-0.312
-0.062	-0.0030	-0.919	0.381	-0.107	0.477	**0.433(a)**	0.943	-0.038	0.571	0.08	-0.246	0.0080	0.023
-0.06	-0.013	-0.971	0.427	-0.124	0.549	0.943	**0.459(a)**	-0.103	0.599	0.046	-0.287	0.0	0.0030
0.015	-0.034	0.033	0.328	0.282	0.158	-0.038	-0.103	**0.899(a)**	-0.249	0.239	-0.158	0.176	-0.078
-0.046	0.0090	-0.535	0.054	0.207	0.055	0.571	0.599	-0.249	**0.449(a)**	0.013	0.13	-0.103	0.078
0.015	-0.044	-0.052	0.215	-0.021	0.0010	0.08	0.046	0.239	0.013	**0.873(a)**	0.23	-0.123	-0.134

0.0080	-0.026	0.371	0.047	0.463	-0.454	-0.246	-0.287	-0.158	0.13	0.23	**0.778(a)**	0.249	-0.267
0.014	-0.012	0.054	-0.155	0.386	-0.221	0.0080	0.0	0.176	-0.103	-0.123	0.249	**0.667(a)**	0.388
0.012	0.01	0.018	-0.643	-0.149	-0.312	0.023	0.0030	-0.078	0.078	-0.134	-0.267	0.388	**0.815(a)**

Con el conjunto de variables S/C, Trect, D-carga, CCR-carga, Ni-carga, V-carga, V50-carga, CarAro-carga, Penet-carga, FE-carga, N2b-carga, SatCarga, ResCarga, C3, T50

Principal Components Attribute Transformer

Correlation matrix

1	0.03	0	0.02	-0.01	0	0.02	-0.01	-0.01	0.02	-0.02	0	0	0.07	0.01	
0.03	1	0.06	0.02	0.05	0.07	-0.01	0.09	0	-0.04	0.03	0.01	-0.05	-0.09	0	
0	0.06	1	0.88	0.68	0.63	0.55	0.82	-0.71	0.21	0.38	-0.72	-0.18	0.07	0.16	
0.02	0.02	**0.88**	1	0.59	0.52	0.57	0.59	-0.79	0.34	0.37	-0.75	-0.06	0.1	0.29	
-0.01	0.05	**0.68**	**0.59**	1	0.6	0.43	0.56	-0.67	-0.15	0.37	-0.52	-0.4	0.1	-0.17	
0	0.07	**0.63**	**0.52**	**0.6**	1	0.44	0.4	-0.55	0.15	0.3	-0.33	-0.17	-0.11	0.28	
0.02	-0.01	**0.55**	**0.57**	**0.43**	**0.44**	1	0.05	-0.59	0.21	0.26	-0.52	0.02	-0.03	0.23	
-0.01	0.09	**0.82**	**0.59**	**0.56**	**0.4**	0.05	1	-0.38	-0.11	0.2	-0.41	-0.37	0.07	-0.11	
-0.01	0	**-0.71**	**-0.79**	**-0.67**	**-0.55**	**-0.59**	**-0.38**	1	-0.24	-0.55	0.76	-0.08	-0.09	-0.21	
0.02	-0.04	0.21	**0.34**	-0.15	0.15	0.21	-0.11	-0.24	1	0.25	-0.49	0.54	0.07	0.63	
-0.02	0.03	**0.38**	**0.37**	**0.37**	**0.3**	0.26	0.2	**-0.55**	0.25	1	-0.57	0.23	0.02	0.06	
0	0.01	**-0.72**	**-0.75**	**-0.52**	**-0.33**	**-0.52**	**-0.41**	**0.76**	**-0.49**	**-0.57**	1	-0.28	-0.14	-0.22	
0	-0.05	-0.18	-0.06	**-0.4**	-0.17	0.02	**-0.37**	-0.08	**0.54**	0.23	-0.28	1	0.01	0.54	
0.07	-0.09	0.07	0.1	0.1	-0.11	-0.03	0.07	-0.09	0.07	0.02	-0.14	0.01	1	-0.04	
0.01	0	0.16	0.29	-0.17	0.28	0.23	-0.11	-0.21	**0.63**	0.06	-0.22	**0.54**	-0.04	1	

Sugerencia:

Se recomienda trabajar con las variables de los componentes cuyos eigenvalues sean superiores a la unidad.

eigenvalue	proportion	cumulative	
5.48929	0.36595	0.36595	0.003S/C+0.019Trect+0.395D-carga+0.387CCR-carga+0.318Ni-carga+0.293V-carga+0.276V50-carga+0.263CarAro-carga-0.376Penet-carga+0.141PIE-carga+0.24 N2b-carga-0.358SatCarga-0.009ResCarga+0.037C3+0.11 T50
2.58279	0.17219	0.53814	-0.012S/C+0.061Trect+0.114D-carga-0.011CCR-carga+0.296Ni-carga+0.074V-carga-0.082V50-carga+0.307CarAro-carga+0.043Penet-carga-0.489PIE-carga-0.113N2b-carga+0.165SatCarga-0.545ResCarga-0.004C3-0.459T50
1.20196	0.08013	0.61827	-0.139S/C+0.394Trect+0.024D-carga-0.015CCR-carga-0.039Ni-carga+0.363V-carga+0.147V50-carga-0.051CarAro-carga+0.061Penet-carga-0.009PIE-carga-0.154N2b-carga+0.206SatCarga-0.082ResCarga-0.721C3+0.266T50
1.04954	0.06997	0.68824	0.768S/C+0.384Trect+0.098D-carga+0.093CCR-carga-0.089Ni-carga+0.047V-carga-0.105V50-carga+0.163CarAro-carga+0.101Penet-carga+0.112PIE-carga-0.273N2b-carga+0.075SatCarga-0.069ResCarga+0.226C3+0.195T50
0.99634	0.06642	0.75466	0.078S/C-0.696Trect+0.059D-carga+0.106CCR-carga-0.01Ni-carga+0.208V-carga+0.267V50-carga-0.084CarAro-carga+0.037Penet-carga+0.027PIE-carga-0.49N2b-carga+0.175SatCarga-0.203ResCarga+0.039C3+0.239T50

0.95594 0.06373 0.81839 0.541S/C-0.098Trect-0.202D-carga-0.14CCR-carga+0.157Ni-carga+0.038V-carga+0.389V50-carga-0.44CarAro-carga-0.182Penet-carga-0.197PIE-carga+0.298N2b-carga-0.014SatCarga+0.024ResCarga-0.213C3-0.243T50

0.7848 0.05232 0.87071 -0.301S/C+0.441Trect-0.1D-carga+0.008CCR-carga+0.11 Ni-carga+0.005V-carga+0.457V50-carga-0.367CarAro-carga-0.086Penet-carga-0.067PIE-carga-0.259N2b-carga+0.062SatCarga-0.094ResCarga+0.505C3+0.043T50

0.61931 0.04129 0.912 0.022S/C+0.055Trect+0.159D-carga+0.242CCR-carga-0.198Ni-carga-0.628V-carga+0.335V50-carga+0.078CarAro-carga+0.029Penet-carga+0.021PIE-carga-0.343N2b-carga-0.292SatCarga+0.006ResCarga-0.314C3-0.241T50

0.39171 0.02611 0.93811 0.017S/C-0.015Trect+0.004D-carga+0.017CCR-carga+0.047Ni-carga-0.097V-carga-0.054V50-carga+0.154CarAro-carga-0.301Penet-carga-0.697PIE-carga-0.176N2b-carga+0.059SatCarga+0.511ResCarga+0.004C3+0.298T50

0.3151 0.02101 0.95912 -0.011S/C-0.025Trect+0.248D-carga-0.027CCR-carga-0.476Ni-carga-0.002V-carga+0.443V50-carga+0.23 CarAro-carga+0.336Penet-carga-0.228PIE-carga+0.449N2b-carga+0.239SatCarga-0.036ResCarga+0.147C3+0.117T50

Eigenvectors

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	
0.0026	-0.0123	-0.1387	**0.7678**	0.0782	0.5409	-0.3014	0.0223	0.0174	-0.0108	S/C
0.0191	0.0611	**0.3939**	0.3844	-0.696	-0.0984	0.4409	0.0553	-0.0153	-0.0247	Trect
0.3954	0.1142	0.0237	0.0982	0.059	-0.2019	-0.1003	0.1593	0.0042	0.2484	D-carga
0.3875	-0.0108	-0.0147	0.0928	0.1055	-0.1402	0.0085	0.2415	0.0172	-0.0271	CCR-carga
0.3184	0.2961	-0.0393	-0.0887	-0.0095	0.1569	0.1101	-0.1978	0.0469	-0.4764	Ni-carga
0.2928	0.0744	**0.3634**	0.0474	0.2078	0.038	0.0052	-0.6281	-0.097	-0.0024	V-carga
0.2761	-0.0822	0.1474	-0.1049	0.2675	0.3894	0.4566	0.3351	-0.0544	0.4427	V50-carga
0.2635	0.307	-0.0514	0.1634	-0.0844	-0.4403	-0.3673	0.0782	0.1539	0.2305	CarAro-carga
-0.3762	0.0425	0.061	0.1008	0.0369	-0.1818	-0.0856	0.0287	-0.3009	0.3361	Penet-carga
0.1407	-0.489	-0.0092	0.1115	0.0274	-0.1968	-0.0673	0.0209	-0.6971	-0.2281	PIE-carga
0.2395	-0.1128	-0.154	-0.2725	-0.4901	0.2978	-0.2593	-0.3429	-0.1764	0.4494	N2b-carga
-0.358	0.1647	0.2061	0.0753	0.1753	-0.0138	0.0624	-0.2915	0.0591	0.2391	SatCarga
-0.0093	**-0.5448**	-0.0817	-0.0689	-0.2031	0.0242	-0.0945	0.0064	0.511	-0.0358	ResCarga
0.0374	-0.0043	-0.7207	**0.2261**	0.0391	-0.2129	0.5053	-0.3143	0.0045	0.1471	C3
0.1099	**-0.4594**	0.266	0.1954	0.2388	-0.2432	0.0433	-0.2409	0.2977	0.1171	T50

EL DETERMINANTE DE LA MATRIZ ES:

1.491505882388929E-6

COEFICIENTE DE ESFERICIDAD DE BARTLETT

2669.729133522495

KMO

0.6108394013112496

MATRIZ ANTI-IMAGEN DE CORRELACIONES

S/C	Trect	D-carga	CCR-carga	Ni-carga	V-carga	V50-carga	CarAro-carga	Penet-carga	PIE-carga	N2b-carga	SatCarga	ResCarga	C3	T50

0.164(a)	-0.04	0.078	-0.066	0.032	-0.059	-0.078	-0.074	0.014	-0.052	0.015	0.01	-0.0010	-0.085	0.024
-0.04	**0.702(a)**	-0.014	0.014	-0.025	0.0	0.012	0.0	-0.031	0.017	-0.042	-0.021	-0.0020	0.084	-0.027
0.078	-0.014	**0.553(a)**	-0.666	0.15	-0.657	-0.922	-0.972	0.03	-0.519	-0.047	0.36	0.012	-0.15	0.066
-0.066	0.014	-0.666	**0.69(a)**	-0.028	0.499	0.527	0.563	0.349	0.214	0.12	-0.091	0.24	0.082	-0.259
0.032	-0.025	0.15	-0.028	**0.796(a)**	-0.415	-0.136	-0.15	0.265	0.185	-0.042	0.415	0.389	-0.169	0.055
-0.059	0.0	-0.657	0.499	-0.415	**0.517(a)**	0.533	0.588	0.136	0.163	-0.068	-0.471	0.033	0.245	-0.257
-0.078	0.012	-0.922	0.527	-0.136	0.533	**0.395(a)**	0.945	-0.031	0.557	0.075	-0.219	0.044	0.169	-0.094
-0.074	0.0	-0.972	0.563	-0.15	0.588	0.945	**0.404(a)**	-0.097	0.578	0.043	-0.271	0.032	0.145	-0.07
0.014	-0.031	0.03	0.349	0.265	0.136	-0.031	-0.097	**0.892(a)**	-0.233	0.232	-0.185	0.188	0.033	0.013
-0.052	0.017	-0.519	0.214	0.185	0.163	0.557	0.578	-0.233	**0.509(a)**	-0.039	0.237	0.059	-0.0040	-0.353
0.015	-0.042	-0.047	0.12	-0.042	-0.068	0.075	0.043	0.232	-0.039	**0.89(a)**	0.14	-0.153	0.066	0.174
0.01	-0.021	0.36	-0.091	0.415	-0.471	-0.219	-0.271	-0.185	0.237	0.14	**0.75(a)**	0.467	-0.057	-0.271
-0.0010	-0.0020	0.012	0.24	0.389	0.033	0.044	0.032	0.188	0.059	-0.153	0.467	**0.592(a)**	-0.011	-0.499
-0.085	0.084	-0.15	0.082	-0.169	0.245	0.169	0.145	0.033	-0.0040	0.066	-0.057	-0.011	**0.32(a)**	0.0
0.024	-0.027	0.066	-0.259	0.055	-0.257	-0.094	-0.07	0.013	-0.353	0.174	-0.271	-0.499	0.0	**0.626(a)**

Con las variables S/C, Trect, D-carga, CCR-carga, Ni-carga, V-carga, V50-carga, CarAro-carga, Penet-carga, PIE-carga, N2b-carga, SatCarga, ResCarga, C3, T10

Principal Components Attribute Transformer

Correlation matrix

1	0.03	0	0.02	-0.01	0	0.02	-0.01	-0.01	0.02	-0.02	0	0	0.07	0.01
0.03	1	0.06	0.02	0.05	0.07	-0.01	0.09	0	-0.04	0.03	0.01	-0.05	-0.09	-0.04
0	0.06	1	0.88	0.68	0.63	0.55	0.82	-0.71	0.21	0.38	-0.72	-0.18	0.07	0.22
0.02	0.02	**0.88**	1	0.59	0.52	0.57	0.59	-0.79	0.34	0.37	-0.75	-0.06	0.1	0.37
-0.01	0.05	**0.68**	**0.59**	1	0.6	0.43	0.56	-0.67	-0.15	0.37	-0.52	-0.4	0.1	-0.15
0	0.07	**0.63**	**0.52**	**0.6**	1	0.44	0.4	-0.55	0.15	0.3	-0.33	-0.17	-0.11	0.17
0.02	-0.01	**0.55**	**0.57**	**0.43**	**0.44**	1	0.05	-0.59	0.21	0.26	-0.52	0.02	-0.03	0.24
-0.01	0.09	**0.82**	**0.59**	**0.56**	**0.4**	0.05	1	-0.38	-0.11	0.2	-0.41	-0.37	0.07	-0.12
-0.01	0	**-0.71**	**-0.79**	**-0.67**	**-0.55**	**-0.59**	**-0.38**	1	-0.24	-0.55	0.76	-0.08	-0.09	-0.31
0.02	-0.04	0.21	**0.34**	-0.15	0.15	0.21	-0.11	-0.24	1	0.25	-0.49	0.54	0.07	0.98
-0.02	0.03	**0.38**	**0.37**	**0.37**	**0.3**	0.26	0.2	**-0.55**	0.25	1	-0.57	0.23	0.02	0.29
0	0.01	**-0.72**	**-0.75**	**-0.52**	**-0.33**	**-0.52**	**-0.41**	**0.76**	**-0.49**	**-0.57**	1	-0.28	-0.14	-0.52
0	-0.05	-0.18	-0.06	**-0.4**	-0.17	0.02	**-0.37**	-0.08	**0.54**	0.23	-0.28	1	0.01	0.62
0.07	-0.09	0.07	0.1	0.1	-0.11	-0.03	0.07	-0.09	0.07	0.02	-0.14	0.01	1	0.06
0.01	-0.04	0.22	**0.37**	-0.15	0.17	0.24	-0.12	**-0.31**	**0.98**	0.29	**-0.52**	**0.62**	0.06	1

Sugerencia:

Se recomienda trabajar con las variables de los componentes cuyos eigenvalues sean superiores a la unidad.

eigenvalue	proportion	cumulative	
5.572	0.37147	0.37147	0.003S/C+0.016Trect+0.388D-carga+0.384CCR-carga+0.305Ni-carga+0.283V-carga+0.273V50-carga+0.251CarAro-carga-0.373Penet-carga+0.164PIE-carga+0.244N2b-carga-0.365SatCarga+0.008ResCarga+0.04 C3+0.177T10
2.86537	0.19102	0.56249	0.011S/C-0.065Trect-0.139D-carga-0.025CCR-carga-0.301Ni-carga-0.127V-carga+0.031V50-carga-0.302CarAro-carga-0.005Penet-carga+0.485PIE-carga+0.113N2b-carga-0.156SatCarga+0.501ResCarga+0.028C3+0.502T10
1.15583	0.07706	0.63955	-0.251S/C+0.394Trect-0.039D-carga-0.064CCR-carga-0.026Ni-carga+0.279V-carga+0.183V50-carga-0.154CarAro-carga-0.003Penet-carga+0.002PIE-carga+0.064N2b-carga+0.102SatCarga+0.033ResCarga-0.79C3+0.02 T10
1.0409	0.06939	0.70894	0.675S/C+0.667Trect+0.065D-carga+0.035CCR-carga-0.08Ni-carga+0.011V-carga-0.165V50-carga+0.159CarAro-carga+0.09 Penet-carga+0.106PIE-carga-0.072N2b-carga+0.022SatCarga-0.014ResCarga+0.048C3+0.085T10
0.98097	0.0654	0.77434	-0.622S/C+0.366Trect+0.108D-carga+0.024CCR-carga-0.076Ni-carga-0.162V-carga-0.469V50-carga+0.39 CarAro-carga+0.116Penet-carga+0.09 PIE-carga+0.089N2b-carga-0.086SatCarga+0.052ResCarga+0.134C3+0.077T10
0.89637	0.05976	0.8341	-0.018S/C-0.28Trect+0.218D-carga+0.192CCR-carga-0.188Ni-carga+0.194V-carga-0.034V50-carga+0.244CarAro-carga+0.219Penet-carga+0.308PIE-carga-0.634N2b-carga+0.139SatCarga-0.222ResCarga-0.178C3+0.242T10
0.7801	0.05201	0.8861	0.303S/C-0.426Trect+0.047D-carga-0.064CCR-carga-0.047Ni-carga-0.011V-carga-0.475V50-carga+0.325CarAro-carga+0.041Penet-carga-0.029PIE-carga+0.402N2b-carga-0.06SatCarga+0.113ResCarga-0.453C3-0.026T10
0.59349	0.03957	0.92567	-0.037S/C-0.028Trect-0.146D-carga-0.228CCR-carga+0.181Ni-carga+0.703V-carga-0.275V50-carga-0.133CarAro-carga+0.044Penet-carga+0.185PIE-carga+0.224N2b-carga+0.299SatCarga-0.147ResCarga+0.283C3+0.163T10
0.36977	0.02465	0.95032	-0.002S/C-0.002Trect+0.043D-carga-0.035CCR-carga-0.044Ni-carga+0.349V-carga-0.111V50-carga+0.137CarAro-carga-0.227Penet-carga-0.32PIE-carga-0.374N2b-carga+0.041SatCarga+0.719ResCarga+0.073C3-0.156T10

Eigenvectors

V1	V2	V3	V4	V5	V6	V7	V8	V9	
0.0029	0.0107	-0.2515	**0.6747**	-0.6223	-0.0182	0.3028	-0.0369	-0.0024	S/C
0.0162	-0.0648	0.394	**0.6673**	0.3657	-0.2796	-0.4261	-0.0284	-0.0024	Trect
0.3876	-0.1395	-0.0393	0.0652	0.1085	0.2182	0.0474	-0.1464	0.0429	D-carga
0.3837	-0.0252	-0.0643	0.0351	0.0241	0.1921	-0.0642	-0.228	-0.0345	CCR-carga
0.3048	-0.3011	-0.0258	-0.0804	-0.0762	-0.1876	-0.0469	0.1814	-0.0435	Ni-carga
0.2832	-0.127	0.279	0.0114	-0.1616	0.1936	-0.0107	0.703	0.3486	V-carga
0.273	0.0307	0.1832	-0.1654	-0.4686	-0.0341	-0.4755	-0.2751	-0.1112	V50-carga
0.2507	**-0.3025**	-0.1541	0.1595	0.3896	0.2443	0.3247	-0.1326	0.1372	CarAro-carga
-0.3731	-0.0047	-0.0027	0.09	0.1162	0.2193	0.0411	0.0445	-0.2268	Penet-carga
0.1637	**0.4853**	0.002	0.1055	0.0902	0.3079	-0.0287	0.1851	-0.3203	PIE-carga
0.2444	0.1126	0.0641	-0.0717	0.0886	-0.6339	0.402	0.2242	-0.3736	N2b-carga
-0.3646	-0.1557	0.1023	0.0224	-0.0858	0.1395	-0.0595	0.2986	0.0411	SatCarga
0.0084	**0.5008**	0.0326	-0.0136	0.0516	-0.2223	0.1133	-0.1469	0.7193	ResCarga
0.0402	0.028	**-0.79**	0.0478	0.1339	-0.1775	-0.4532	0.2831	0.073	C3
0.1774	**0.5023**	0.0201	0.0853	0.077	0.2417	-0.0259	0.163	-0.1561	T10

EL DETERMINANTE DE LA MATRIZ ES:

7.769435763650778E-8

COEFICIENTE DE ESFERICIDAD DE BARTLETT

3257.7261567005667

KMO

0.6124598329746017

MATRIZ ANTI-IMAGEN DE CORRELACIONES

S/C	Trect	D-carga	CCR-carga	Ni-carga	V-carga	V50-carga	CarAro-carga	Penet-carga	PIE-carga	N2b-carga	SatCarga	ResCarga	C3	T10
0.105(a)	-0.041	0.088	-0.081	0.05	-0.071	-0.086	-0.083	0.041	-0.096	0.0090	0.015	-0.034	-0.086	0.087
-0.041	**0.711(a)**	-0.015	0.011	-0.027	-0.0040	0.012	0.0	-0.034	0.018	-0.037	-0.029	-0.0070	0.084	-0.017
0.088	-0.015	**0.561(a)**	-0.682	0.177	-0.673	-0.923	-0.972	0.077	-0.294	-0.061	0.385	-0.032	-0.151	0.15
-0.081	0.011	-0.682	**0.689(a)**	-0.069	0.49	0.533	0.576	0.259	0.263	0.173	-0.162	0.231	0.086	-0.235
0.05	-0.027	0.177	-0.069	**0.779(a)**	-0.445	-0.155	-0.172	0.321	-0.166	-0.055	0.428	0.282	-0.168	0.237
-0.071	-0.0040	-0.673	0.49	-0.445	**0.493(a)**	0.539	0.601	0.065	0.225	-0.02	-0.563	0.012	0.251	-0.211
-0.086	0.012	-0.923	0.533	-0.155	0.539	**0.413(a)**	0.946	-0.067	0.274	0.095	-0.25	0.059	0.171	-0.119
-0.083	0.0	-0.972	0.576	-0.172	0.601	0.946	**0.418(a)**	-0.133	0.293	0.058	-0.296	0.064	0.146	-0.131
0.041	-0.034	0.077	0.259	0.321	0.065	-0.067	-0.133	**0.864(a)**	-0.377	0.215	-0.186	0.016	0.026	0.325
-0.096	0.018	-0.294	0.263	-0.166	0.225	0.274	0.293	-0.377	**0.528(a)**	0.022	0.069	0.453	0.012	-0.958
0.0090	-0.037	-0.061	0.173	-0.055	-0.02	0.095	0.058	0.215	0.022	**0.907(a)**	0.199	-0.059	0.067	-0.015
0.015	-0.029	0.385	-0.162	0.428	-0.563	-0.25	-0.296	-0.186	0.069	0.199	**0.766(a)**	0.355	-0.059	-0.026
-0.034	-0.0070	-0.032	0.231	0.282	0.012	0.059	0.064	0.016	0.453	-0.059	0.355	**0.614(a)**	-0.0040	-0.51
-0.086	0.084	-0.151	0.086	-0.168	0.251	0.171	0.146	0.026	0.012	0.067	-0.059	-0.0040	**0.32(a)**	-0.014
0.087	-0.017	0.15	-0.235	0.237	-0.211	-0.119	-0.131	0.325	-0.958	-0.015	-0.026	-0.51	-0.014	**0.583(a)**

Anexo E. Gams

La herramienta computacional GAMS (*General Algebraic Modeling System*) es un poderoso paquete matemático que permite entre muchas opciones, el modelamiento de sistemas lineales, no lineales y mixtos, de programación entera, y problemas de optimización.

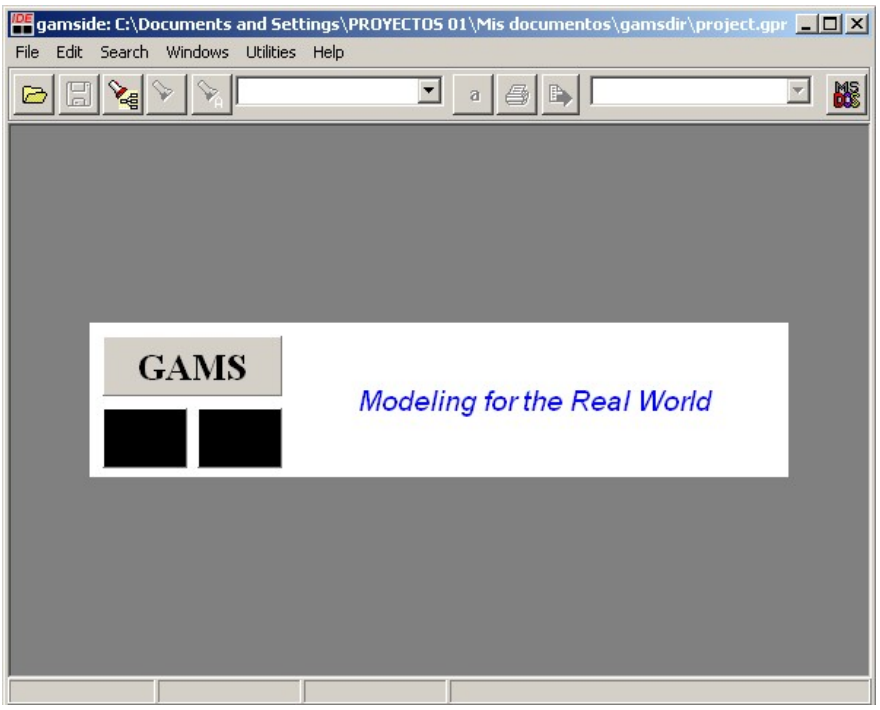
Este paquete ha sido diseñado para trabajar problemas de gran magnitud, y para ser usado desde computadoras personales, hasta *mainframes* y supercomputadoras. GAMS permite al usuario modelar una situación de una manera sencilla, mediante un "SETUP" de opciones, disminuyendo los tiempos de respuesta para cada problema. Disponer de diferentes paquetes de solución o "SOLVERS", tiene la capacidad de manipular grandes problemas, y transformarlos de su manera lineal a no lineal sin mucha dificultad.

La versión de GAMS que se usó en este proyecto es:

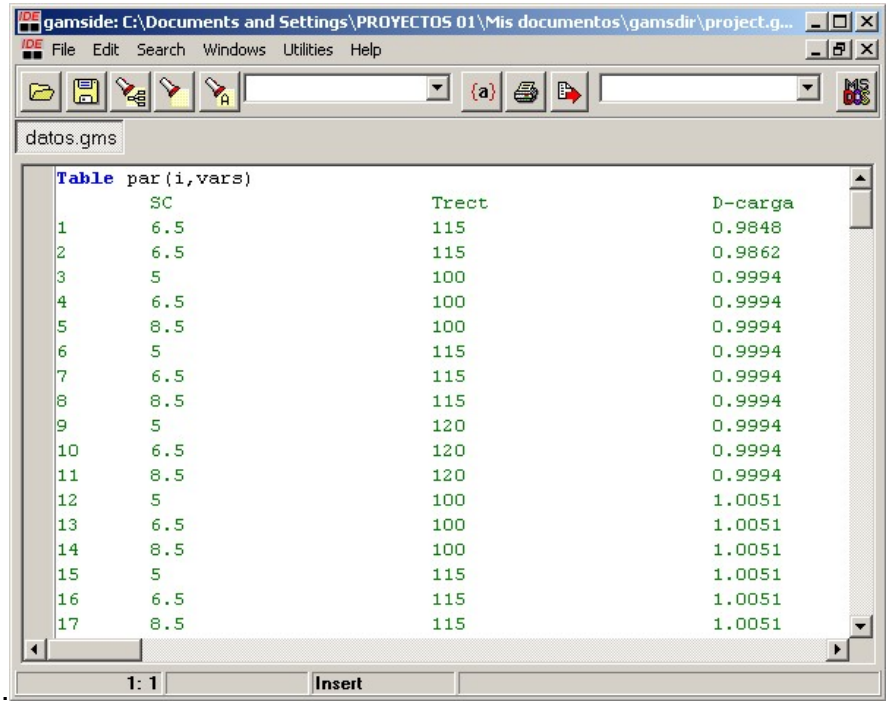
GAMS IDE	2.0.28.0
Module	GAMS Rev 140
Lic date	Nov 11, 2004
Build	VIS 21.5 140

Una vista a GAMS

Ventana Principal

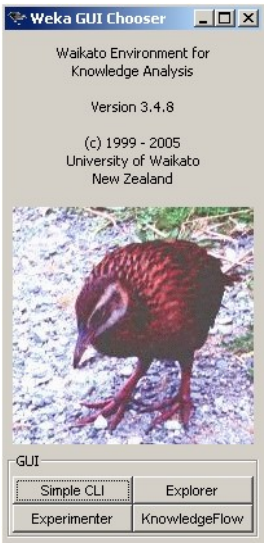


Datos Cargados

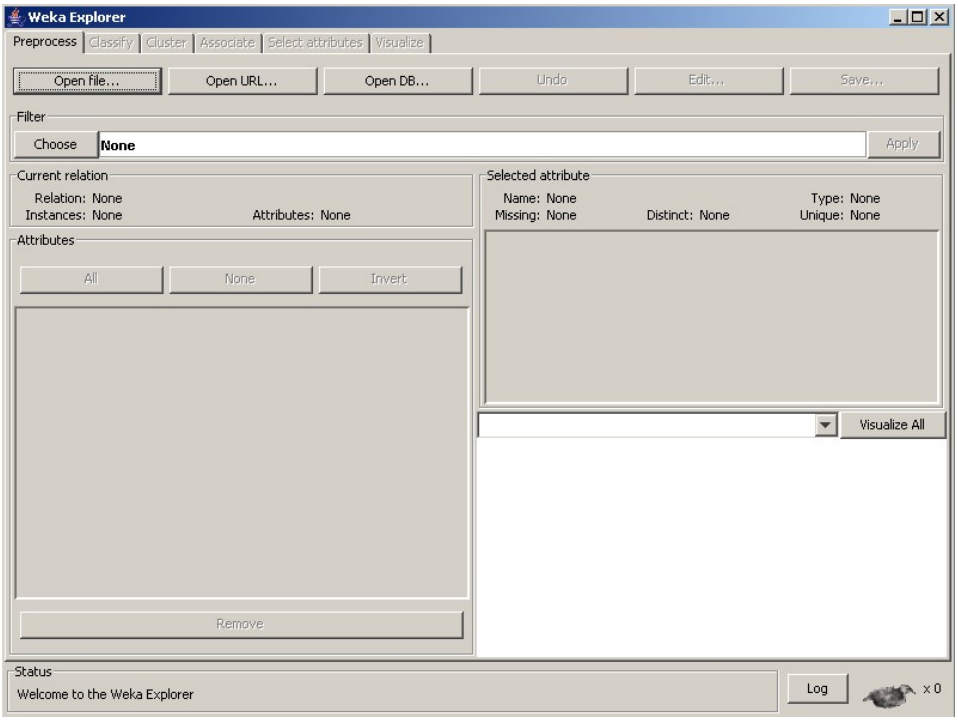


Anexo F. *Weka*

Vista a *WEKA*



Opción Explorer



Anexo G. Manual de usuario

El SPP 2.0 es un prototipo computacional desarrollado bajo el ambiente de Microsoft Windows XP, para trabajar bajo este y programado en su totalidad en la plataforma Java y el entorno de desarrollo NETBEANS de Sun Microsystems; esta aplicación posee un módulo de comunicaciones a la base de datos SILAB del ICP, protocolizada con Oracle. La función de este prototipo es emplear técnicas de Minería de Datos y técnicas estadísticas para estimar rendimientos y calidades de productos del proceso de extracción del DMO.

Este manual encierra todo el manejo del Sistema de Predicción y Propiedades (SPP 2.0) mediante algoritmos de Minería de Datos y estadísticos, y describe todas las funciones y elementos que lo conforman.

Los módulos que conforman este sistema son:

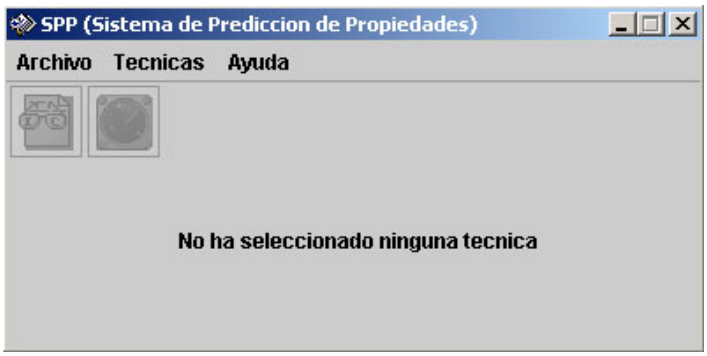
- **Manejo de Información:** contiene todo lo referente a la manipulación de archivos, así como cargar y/o agregar datos de un archivo de datos o de la base de datos. Además se pueden guardar datos previamente cargados y los resultados de los análisis resueltos por el prototipo. Los datos pueden ser cargados y manipulados para hacer posible el análisis, presentarlos al usuario y guardar la información.

- **Análisis:** se realizan los análisis respectivos según la técnica de Minería de Datos o estadística empleada, así como Análisis de Regresión, Árboles de Decisión, Cluster, Componentes Principales, Redes Bayesianas y Análisis de Varianza.

Manual del usuario de SPP

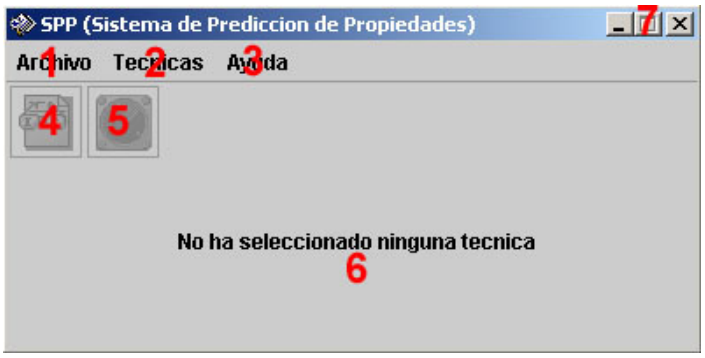
El SPP (Sistema de Predicción de Propiedades) fue diseñado y mejorado con un entorno más amable y fácil de usar; de esta manera los usuarios podrán navegar a través de él sin inconvenientes y con todos los soportes de ayuda necesarios para su buen uso y desempeño. A continuación se presenta una descripción completa y clara del sistema como tal y cada una de sus funciones.
Inicialmente se identifica la pantalla del SPP como se aprecia en la Figura 1.

Figura *1. Ventana inicial del SPP. Fuente Autores.*



En esta ventana se pueden identificar varios objetos, que serán descritos a continuación (ver Figura 2):

Figura *2. Ventana inicial del SPP (con funciones). Fuente Autores.*

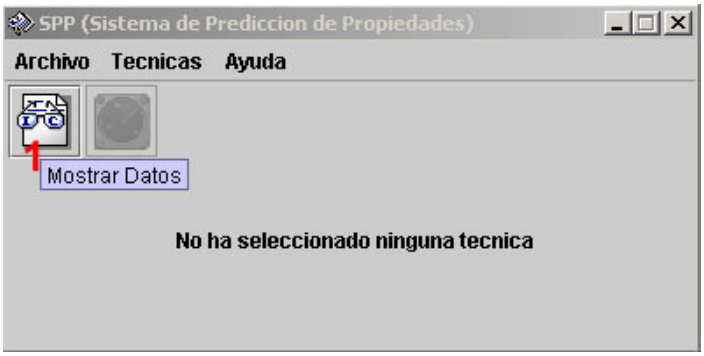


1. Menú Archivo: en éste menú se encuentran las funciones básicas de la aplicación como: Cargar Datos, Agregar Datos, Guardar, y Salir.
2. Menú Técnicas: en éste menú se encuentran disponibles las diferentes Técnicas de Minería que se pueden aplicar en el prototipo como: Componentes principales, Análisis de Regresión, Análisis de Varianza, Cluster, Árboles de Decisión, Redes Bayesianas, GAMS.
3. Menú Ayuda: aquí se encuentra el documento de ayuda, y uno adicional de ayuda sobre la metodología y recomendaciones de uso para la serie de pasos en el proceso de descubrimiento del conocimiento.
4. Botón ver datos: éste botón permite visualizar en una Tabla los datos que se han seleccionado de un archivo, o de la base de datos del SILAB.
5. Botón analizar: éste botón inicia los procedimientos de las técnicas seleccionadas.
6. Mensajes: en ésta sección salen mensajes sobre las operaciones que está procesando la aplicación para guiar al usuario.

7. Botones básicos de ventanas de Windows: son los botones de control de la ventana, los cuales permiten minimizar, maximizar, o cerrar la ventana respectivamente.

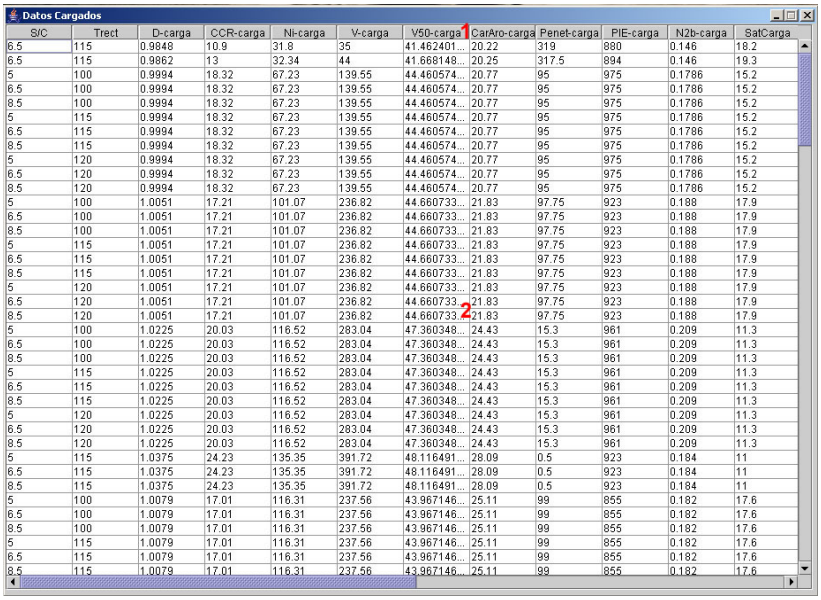
El botón ver datos que se encuentra en la pantalla principal de la aplicación permite ver la sábana que se ha cargado con anterioridad (referirse a la Figura 6) como se aprecia en la Figura 3.

Figura *3. Ventana inicial del SPP. Botón mostrar datos. Fuente Autores.*



1. Botón Mostrar datos: muestra los datos cargados con anterioridad en una Tabla como la que se ven en la Figura 4.

Figura *4. Tabla de datos ya cargados. Fuente Autores.*



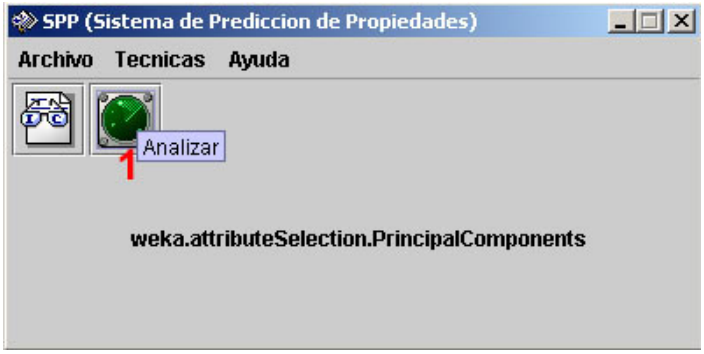
1. En esta sección se encuentran los nombres de todas las variables (dependientes e independientes).

2. En esta sección se encuentran los valores de las variables cargadas.

El botón analizar inicia los procesos seleccionados sobre los datos cargadazos al ser oprimido, como se puede ver en la Figura 5.

Figura *5. Ventana inicial del SPP. Botón Analizar datos. Fuente Autores.*

.

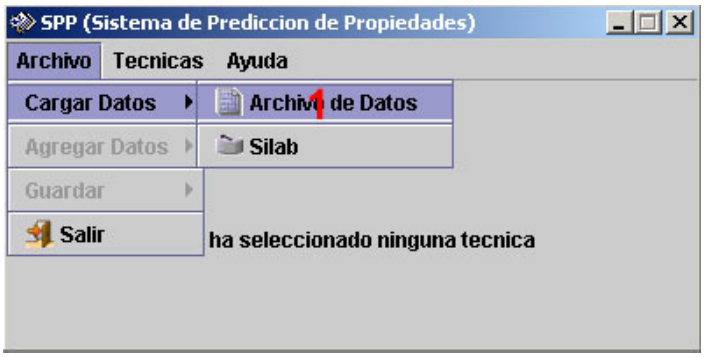


1. Botón Analizar: este botón inicia los procesos seleccionados sobre los datos, para llevarlos a cabo, es necesario hacer clic sobre este.

Menú Archivo

En las siguientes dos ventanas se puede apreciar el procedimiento de cargado de datos, ya sea desde archivo, o desde la base de datos SILAB. Desde archivo (Figura 6):

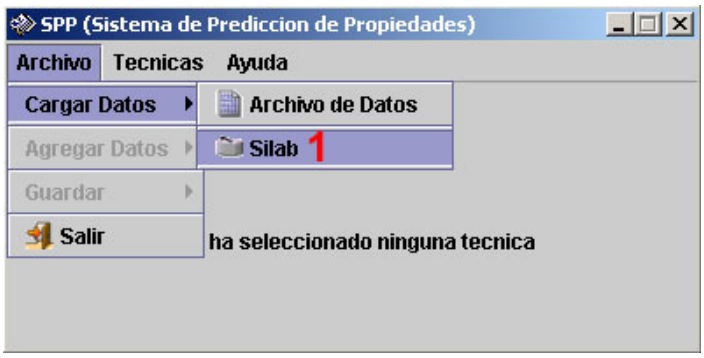
Figura *6. Ventana inicial cargar datos de archivo. Fuente Autores.*



1. Archivo de Datos: permite cargar los datos desde un archivo tipo CSV (texto plano separado por comas (,)).

Desde Base de Datos (Figura 7):

Figura *7. Ventana inicial cargar datos desde el SILAB. Fuente Autores.*



1. SILAB: permite cargar los datos directamente desde la base de datos SILAB, este procedimiento aplica técnicas de *DATAWAREHOUSE* y normalización de los datos.

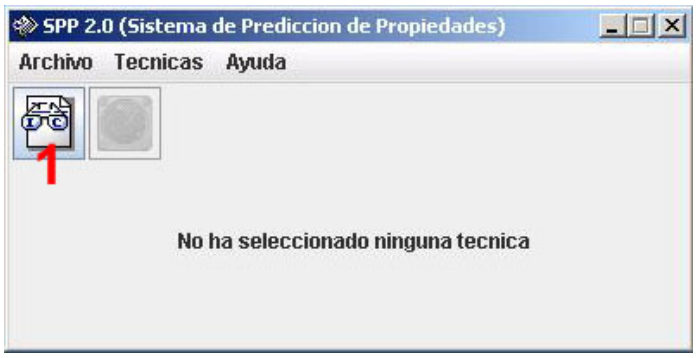
Una vez seleccionado la opción “Silab” se abre una pantalla donde se cargan los SampleID de los elementos a consultar en la base de datos, este archivo es de tipo CSV.

Figura *8: Ventana Identificador de Samples.*

1. Se puede apreciar el archivo de referencias SampleID.csv, que es el archivo que se debe cargar para las consultas a la base de datos. (El archivo debe constar de cuatro columnas Fondo de Vacío, DMO, Demex y Solvente).

Si la consulta a la base de de datos es exitosa, se activará el botón “ver datos” en la ventana principal de la aplicación, como se aprecia en la Figura 9.

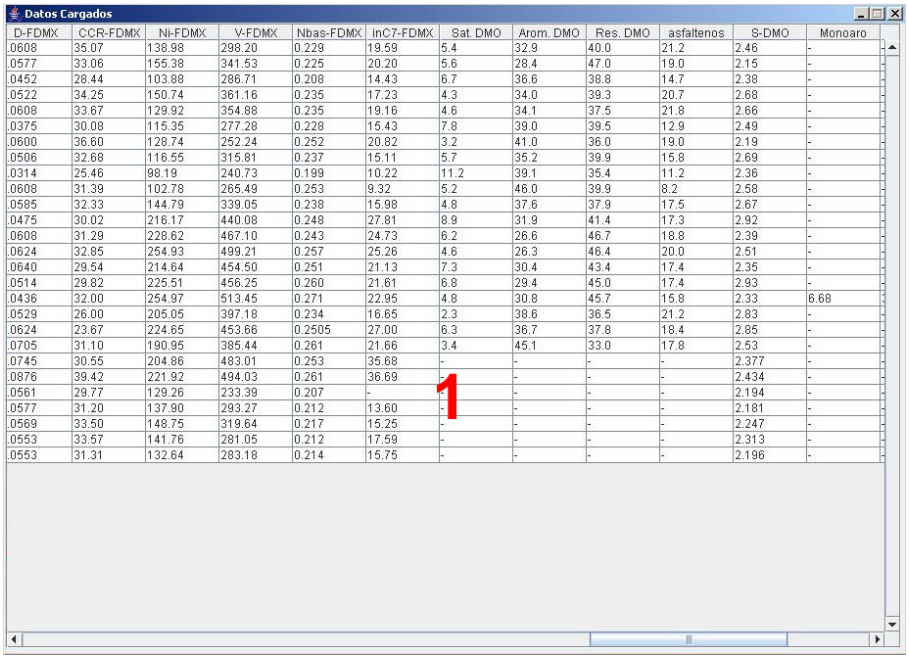
Figura *9: Botón ver datos.*



1. Botón ver datos, después de una consulta exitosa a la base de datos.

Una vez cargados los datos se pueden visualizar y modificar haciendo clic sobre el botón ver datos, se abrirá una ventana con la matriz de datos:

Figura *10: Ventana con datos cargados.*



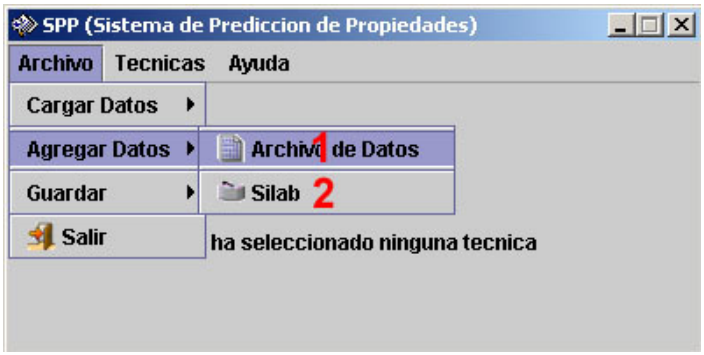
D-FDMX	CCR-FDMX	Ni-FDMX	V-FDMX	Nbas-FDMX	inC7-FDMX	Sat. DMO	Arom. DMO	Res. DMO	asfaltenos	S-DMO	Monoaro
.0608	35.07	138.98	298.20	0.229	19.59	5.4	32.9	40.0	21.2	2.46	-
.0577	33.06	155.38	341.53	0.225	20.20	5.6	28.4	47.0	19.0	2.15	-
.0452	28.44	103.88	286.71	0.208	14.43	6.7	36.6	38.8	14.7	2.38	-
.0522	34.25	150.74	361.16	0.235	17.23	4.3	34.0	39.3	20.7	2.68	-
.0608	33.67	129.92	354.88	0.235	19.16	4.6	34.1	37.5	21.8	2.66	-
.0375	30.08	115.35	277.28	0.228	15.43	7.8	39.0	39.5	12.9	2.49	-
.0600	36.60	128.74	252.24	0.252	20.82	3.2	41.0	36.0	19.0	2.19	-
.0506	32.68	116.55	315.81	0.237	15.11	5.7	35.2	39.9	15.8	2.69	-
.0314	25.46	98.19	240.73	0.199	10.22	11.2	39.1	35.4	11.2	2.36	-
.0608	31.39	102.78	265.49	0.253	9.32	5.2	46.0	39.9	8.2	2.58	-
.0585	32.33	144.79	339.05	0.238	15.98	4.8	37.6	37.9	17.5	2.67	-
.0475	30.02	216.17	440.08	0.248	27.81	8.9	31.9	41.4	17.3	2.92	-
.0608	31.29	228.62	467.10	0.243	24.73	6.2	26.6	46.7	18.8	2.39	-
.0624	32.85	254.93	499.21	0.257	25.26	4.6	26.3	46.4	20.0	2.51	-
.0640	29.54	214.64	454.50	0.251	21.13	7.3	30.4	43.4	17.4	2.35	-
.0514	29.82	225.51	456.25	0.260	21.61	6.8	29.4	45.0	17.4	2.93	-
.0436	32.00	254.97	513.45	0.271	22.95	4.8	30.8	45.7	15.8	2.33	6.68
.0529	26.00	205.05	397.18	0.234	16.65	2.3	38.6	36.5	21.2	2.83	-
.0624	23.67	224.65	453.66	0.2505	27.00	6.3	36.7	37.8	18.4	2.85	-
.0705	31.10	190.95	385.44	0.261	21.66	3.4	45.1	33.0	17.8	2.53	-
.0745	30.55	204.86	483.01	0.253	35.68	-	-	-	-	2.377	-
.0876	39.42	221.92	494.03	0.261	36.69	-	-	-	-	2.434	-
.0561	29.77	129.26	233.39	0.207	-	-	-	-	-	2.194	-
.0577	31.20	137.90	293.27	0.212	13.60	-	-	-	-	2.181	-
.0569	33.50	148.75	319.64	0.217	15.25	-	-	-	-	2.247	-
.0553	33.57	141.76	281.05	0.212	17.59	-	-	-	-	2.313	-
.0553	31.31	132.64	283.18	0.214	15.75	-	-	-	-	2.196	-

1. Datos cargados desde la base de datos Silab.

En caso de que los datos deban ser complementados, el programa contiene la opción Agregar datos, que permite el mismo tratamiento de abrir datos, pero con la

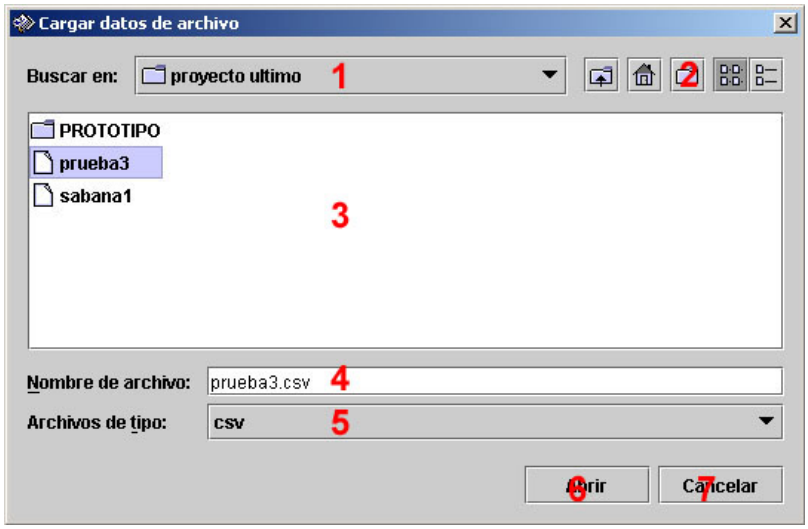
ventaja de agregarlos a la sábana actual, como se puede apreciar en la Figura 11, se puede ver como se agregan datos desde el archivo **(1)** o desde la base de datos del SILAB (**2**).

Figura *11. Ventana inicial agregar datos desde archivo. Fuente Autores.*



Cuando se hace uso de las opciones anteriores de “Cargar Datos”, aparece una ventana similar a la de la Figura 12.

Figura *12. Cargar datos desde archivo. Fuente Autores.*

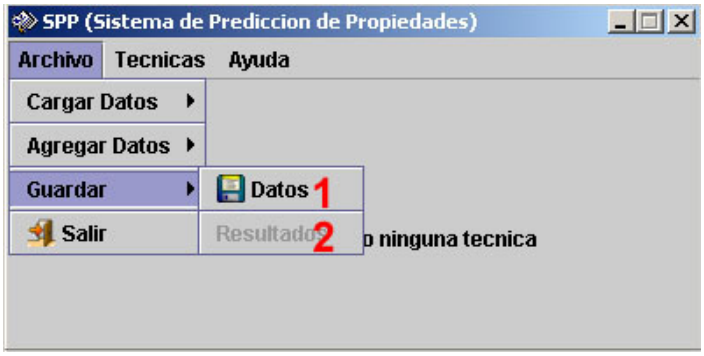


1. Buscar en: esta opción describe la ruta donde reposa el archivo que se quiere abrir, con los *Samples id*.
2. Botones básicos de ventana: aquí se encuentran los botones básicos de las ventanas de Windows, donde están las funciones de navegar por carpetas, y organización de iconos según sus características.
3. Espacio de navegación: aquí se puede ver la lista de archivos, además las carpetas que se encuentran dentro de la carpeta actual de búsqueda.

4. Nombre de archivo: en este campo se ubica el nombre del archivo que se va a abrir.
5. Archivos de tipo: en este campo se especifican qué tipos de archivos se visualizarán en el campo de navegación como válidos, para ser abiertos (en nuestro caso los archivos cuya extensión sea CSV).
6. Botón Abrir: una ves buscado y seleccionado el archivo, se usa el botón abrir para empezar a trabajar con ese archivo.
7. Botón Cancelar: con este botón se cancela la operación de esta ventana.

La opción de guardar los datos esta disponible, con el objetivo de guardar la sábana de datos actual, para futuras interpretaciones u operaciones, como se puede ver en la Figura 13.

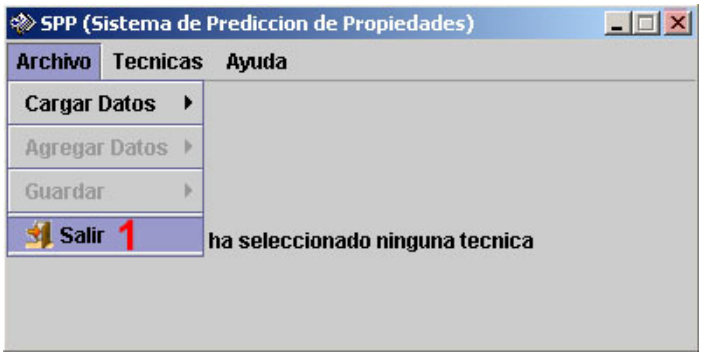
Figura *13. Guardar. Fuente Autores.*



1. Datos: guarda los datos que se encuentran en la aplicación actualmente.
2. Resultados: guarda la ficha de resultados del proceso de las técnicas en archivos de texto plano en formato .RTF.

La ultima opción del menú de Archivo, es salir, y simplemente cumple la función de cerrar la aplicación, similar al icono con la X en la esquina superior derecha, como se aprecia en la Figura 14.

Figura *14. Salir. Fuente Autores.*



1. Salir: control para cerrar la aplicación.

Menú Técnicas

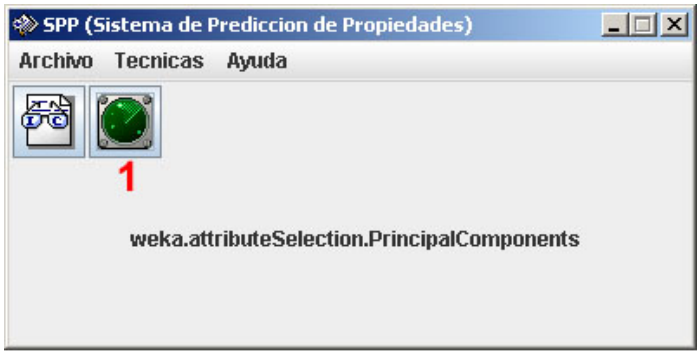
Las siguientes ventanas explican la estructura de las técnicas que el prototipo usa para los procesos de Minería de Datos: El menú de técnicas se puede apreciar en la Figura 15.

Figura *15. Menú de técnicas disponibles. Fuente Autores.*

1. Técnicas: Esta opción despliega las diferentes técnicas que se pueden utilizar para el proceso de Minería de Datos, estas no se encuentran en orden, y su jerarquía esta asociada a los diferentes tipos de procesos; referirse a menú ayuda, metodología.

Cuando se seleccione una de las técnicas el botón “analizar” se activará, y entonces se podrá proceder con las técnicas, como se puede ver en la Figura 16.

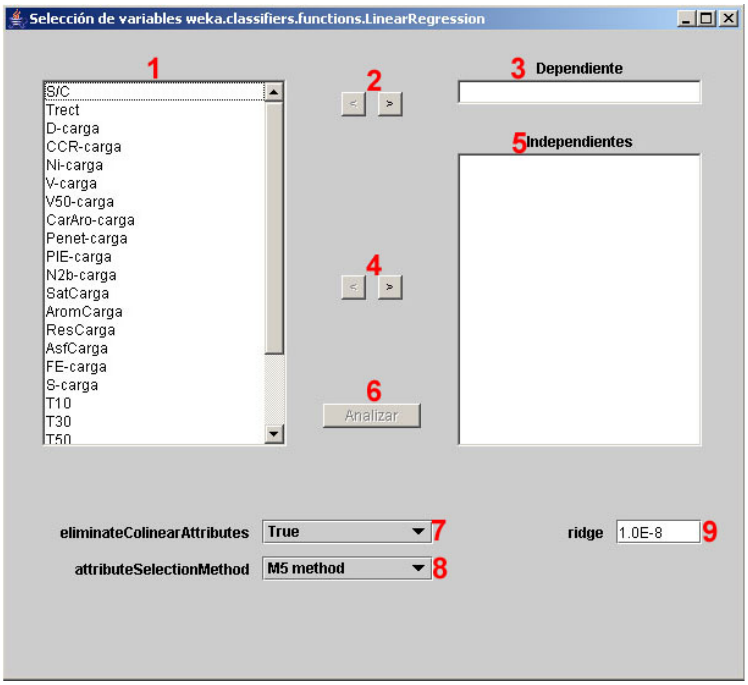
Figura *16: Botón analizar.*



1. **Botón Analizar**: este botón cuando es seleccionado, llama a la ventana de selección de variables y opciones para la ejecución de las técnicas.

2. Análisis de Regresión: esta opción inicia el proceso de análisis de regresión, aplicado a los datos seleccionados previamente (referirse a la Figura 6), seleccionando esta opción se puede ver la ventana de selección de variables para este proceso en la Figura 17.

Figura *17. Ventana Análisis de regresión. Fuente Autores.*

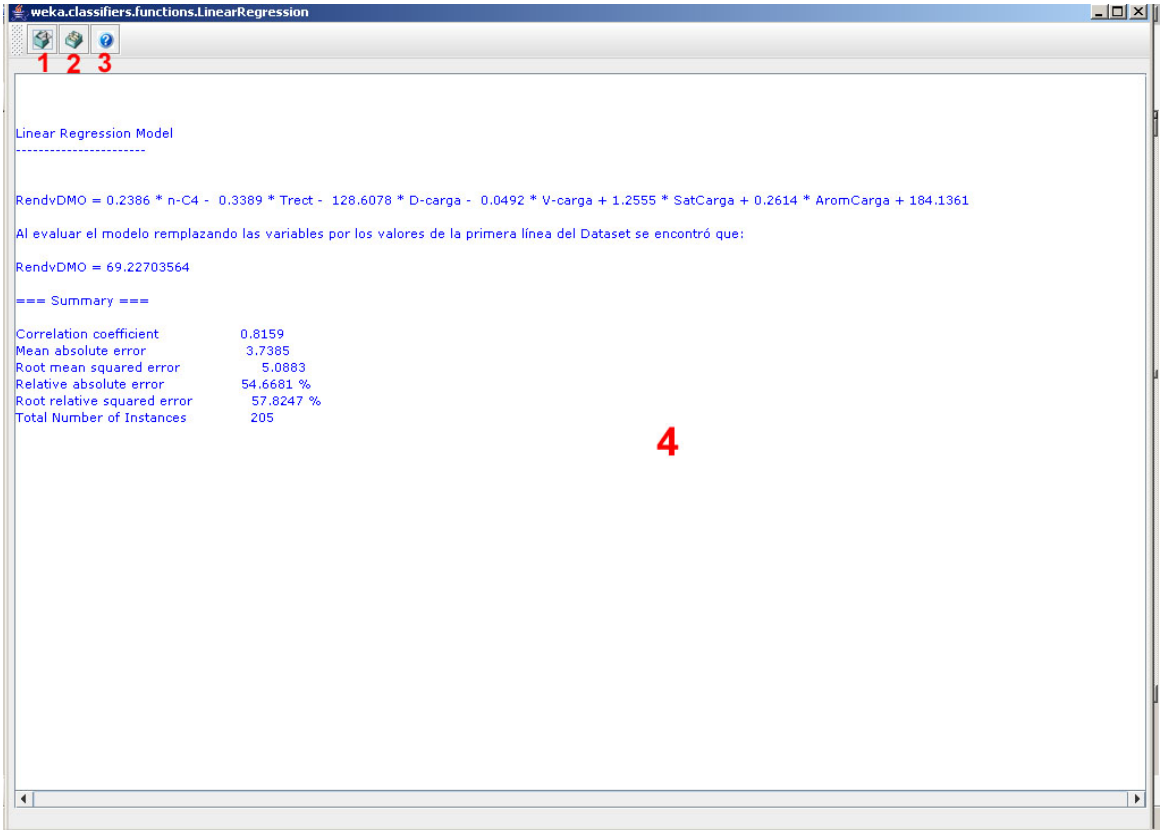


1. En este campo se encuentran todas las variables disponibles cargadas en la sábana de datos, tanto las dependientes como las independientes.
2. Estos botones son los controles de desplazamiento de las variables hacia el campo dependiente, esto significa que en el campo **3** se pondrá la variable dependiente que se encuentra en la lista de variables, si es errado, se puede quitar la selección con botón que tiene el símbolo "<".
3. Campo para ubicar la variable dependiente.
4. Estos botones son los controles de desplazamiento de las variables hacia el campo independiente, esto significa que en el campo **5** se pondrán las variables independientes que se encuentran en la lista de variables, si son errados, se puede quitar la selección con botón que tiene el símbolo "<".
5. Campo para ubicar las variables independientes.
6. Este botón permite iniciar los análisis correspondientes a éste proceso.
7. El *EliminateColinearAttributes* dependiendo del valor que se le halla dado (*True* o Falsee) elimina o no la colinealidad o los puntos que se encuentran sobre la misma recta entre los atributos elegidos.

8. El campo *AttributeSlectionMethod* elige el tipo de algoritmo que se usará para el análisis de regresión, el más óptimo es el que se encuentra por defecto.
9. En el campo *ridge* se mide el "coeficiente de riesgo". Cresta.

Una vez ejecutado este análisis se puede referir a la ventana de resultados Correspondiente a la Figura 18.

Figura *18. Ventana de Resultados de regresión lineal. Fuente Autores.*

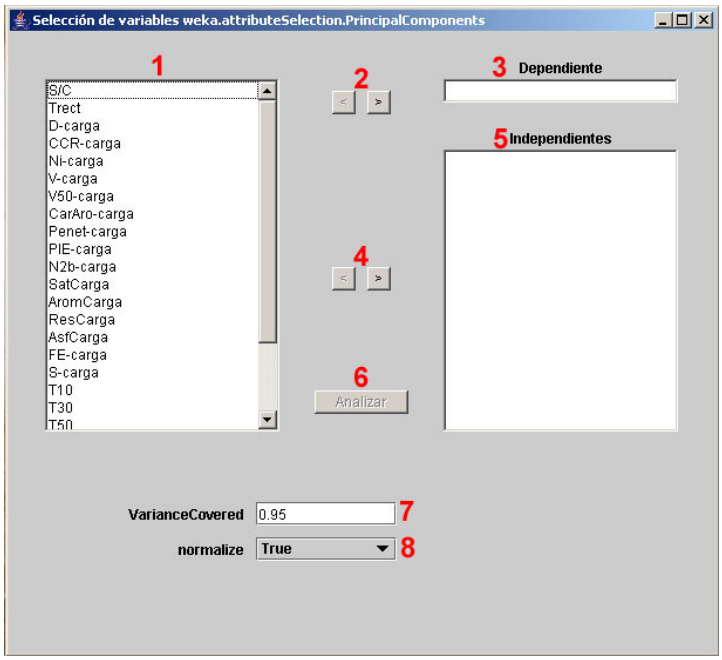


1. Este icono permite guardar los resultados del análisis.
2. Este icono permite acumular resultados de diferentes técnicas en forma contigua, agregando más información dentro del mismo archivo.
3. Este icono permite desplegar la interpretación de resultados de la técnica.
4. Campo donde se muestran los resultados de la técnica.

Las ventanas de resultados y sus funciones (los iconos que se encuentran en la esquina superior izquierda) son iguales en todas las técnicas y funcionan de manera semejante.

3. Componentes principales: ésta opción inicia el proceso de componentes principales, aplicado a los datos seleccionados previamente (referirse a la Figura 6), seleccionando esta opción se puede ver la ventana de selección de variables para este proceso en la Figura 19.

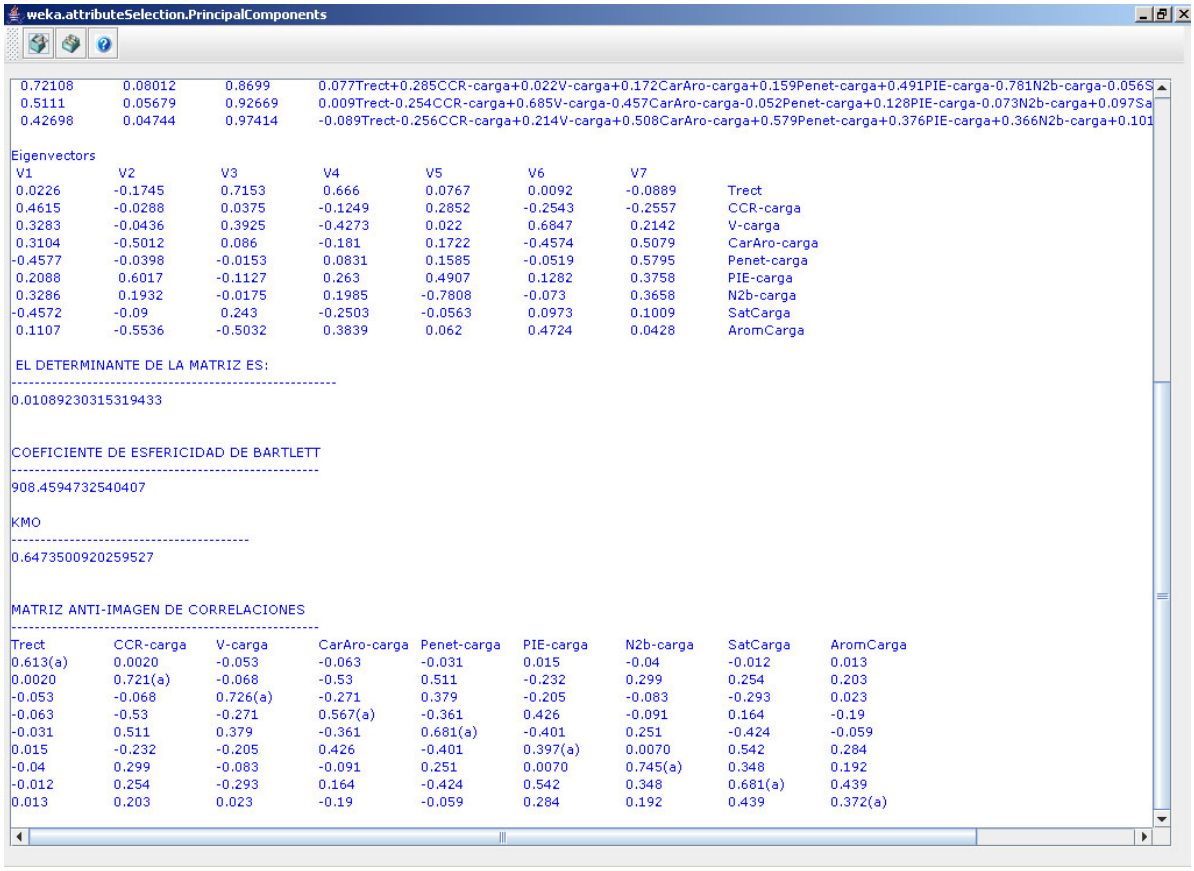
Figura *19. Ventana de Análisis de Componentes Principales. Fuente Autores.*



1. En este campo se encuentran todas las variables disponibles cargadas en la sábana de datos, tanto las dependientes como las independientes.
2. Estos botones son los controles de desplazamiento de las variables hacia el campo dependiente, esto significa que en el campo **3** se pondrá la variable dependiente que se encuentra en la lista de variables, si es errado, se puede quitar la selección con botón que tiene el símbolo “<”.
3. Campo para ubicar la variable dependiente.
4. Estos botones son los controles de desplazamiento de las variables hacia el campo independiente, esto significa que en el campo **5** se pondrán las variables independientes que se encuentran en la lista de variables, si son errados, se puede quitar la selección con botón que tiene el símbolo “<”.
5. Campo para ubicar las variables independientes.
6. Este botón permite iniciar los análisis correspondientes a éste proceso.
7. El campo *VarianceCovered* toma un valor porcentual (generalmente 0.95 o 95%) para explicar los componentes. En ciencias no hay un estándar determinado para este valor, pero una varianza explicada de por lo menos el 95% es aceptable.
8. El campo *Normalize* permite normalizar los datos en la ejecución de la técnica si es requerido.

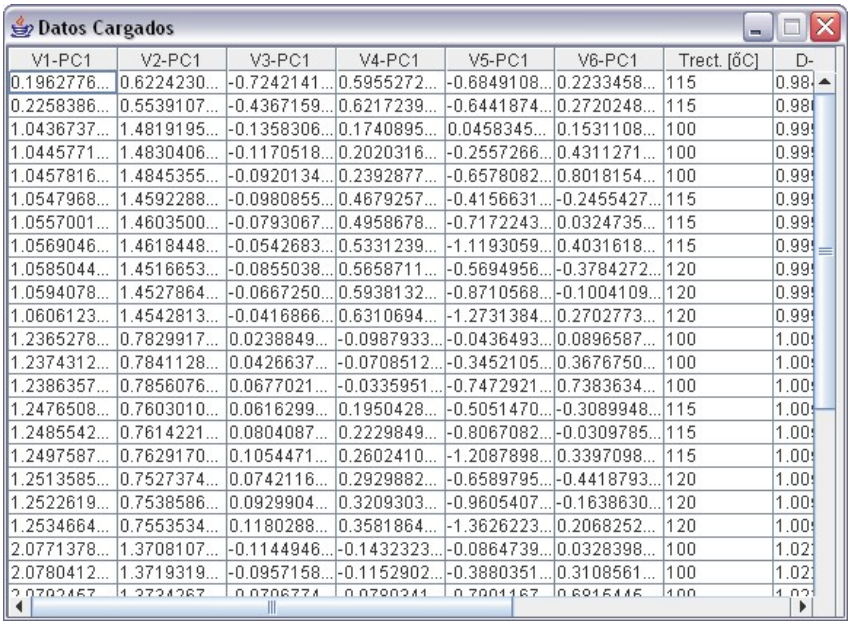
Una vez ejecutado este análisis se puede referir a la ventana de resultados correspondiente a la Figura 20.

Figura *20. Ventana de Resultados de Análisis de Componentes Principales. Fuente Autores.*



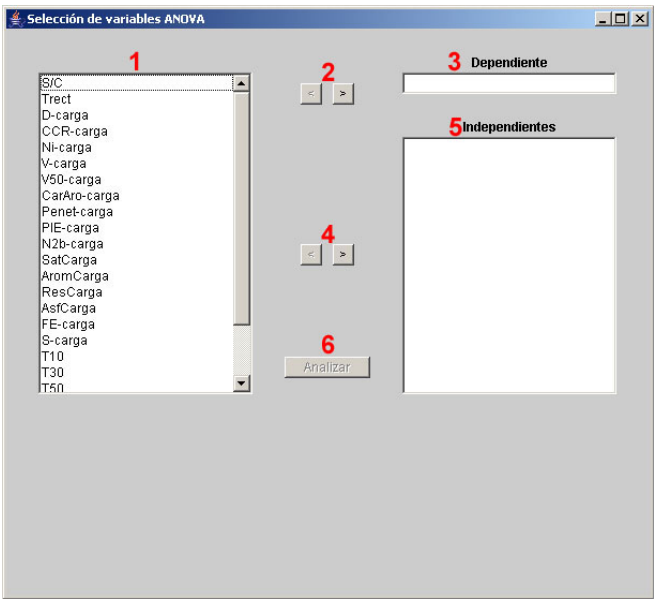
Finalmente esta técnica reduce el número de variables haciendo agrupaciones en componente, estos son considerados las nuevas variables de trabajo, estas variables quedan almacenadas en la sábana de datos que tiene el programa cargada, como se puede apreciar en la Figura 21.

Figura *21. Tabla con las nuevas variables generadas por el análisis de componentes principales Fuente Autores.*



4. Análisis de Varianza "ANOVA": Esta opción inicia el proceso de análisis de Varianza ANOVA, aplicado a los datos seleccionados previamente, seleccionando ANOVA. Una vez ejecutado, seleccionando esta opción se puede ver la ventana de selección de variables para este proceso en la Figura 22.

Figura *22. Ventana Análisis de Varianza ANOVA. Fuente Autores.*



1. En este campo se encuentran todas las variables disponibles cargadas en la sábana de datos, tanto las dependientes como las independientes.
2. Estos botones son los controles de desplazamiento de las variables hacia el campo dependiente, esto significa que en el campo **3** se pondrá la variable dependiente que se encuentra en la lista de variables, si es errado, se puede quitar la selección con botón que tiene el símbolo "<".
3. Campo para ubicar la variable dependiente.
4. Estos botones son los controles de desplazamiento de las variables hacia el campo independiente, esto significa que en el campo **5** se pondrán las variables independientes que se encuentran en la lista de variables, si son errados, se puede quitar la selección con botón que tiene el símbolo "<".
5. Campo para ubicar las variables independientes.
6. Este botón permite iniciar los análisis correspondientes a éste proceso.

Finalmente se obtiene la ventana de resultados para Análisis de Varianza ANOVA, Figura 23.

Figura *23. Ventana de Resultados de Análisis de Varianza ANOVA .Fuente Autores.*

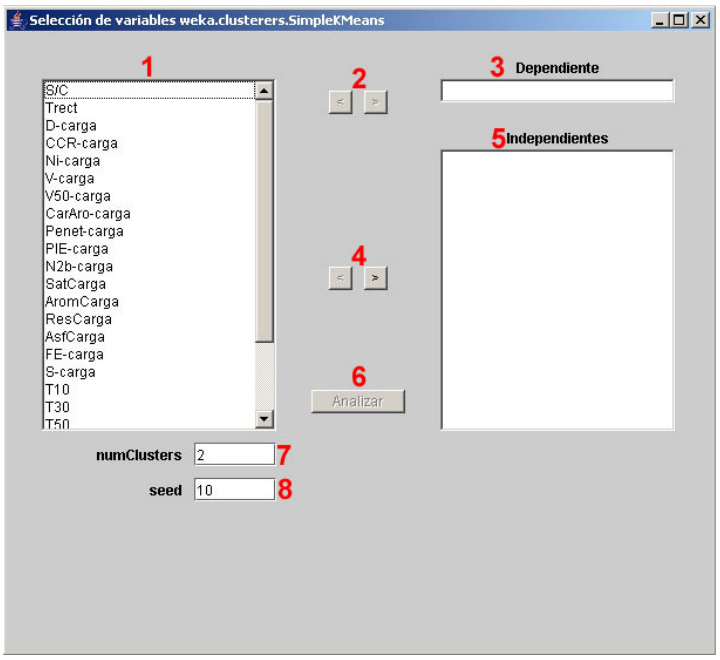
Contrastes individuales de la t
==========

Variables	t Calculado	t Tabla	Resultado
Constante	2.48	1.96	*
Trect	6.51	1.96	*
D-carga	0.40	1.96	-
CCR-carga	2.25	1.96	*
Ni-carga	2.92	1.96	*
V-carga	1.88	1.96	-
V50-carga	0.74	1.96	-
Penet-carga	0.41	1.96	-
FE-carga	1.90	1.96	-
S-carga	0.75	1.96	-
T10	2.50	1.96	*

*: Significativo
-: No significativo

5. Cluster: Esta opción inicia el proceso de *Cluster/ SimpleKMeans,* aplicado a los datos seleccionados previamente. Una vez ejecutado, seleccionando esta opción se puede ver la ventana de selección de variables para este proceso en la Figura 24.

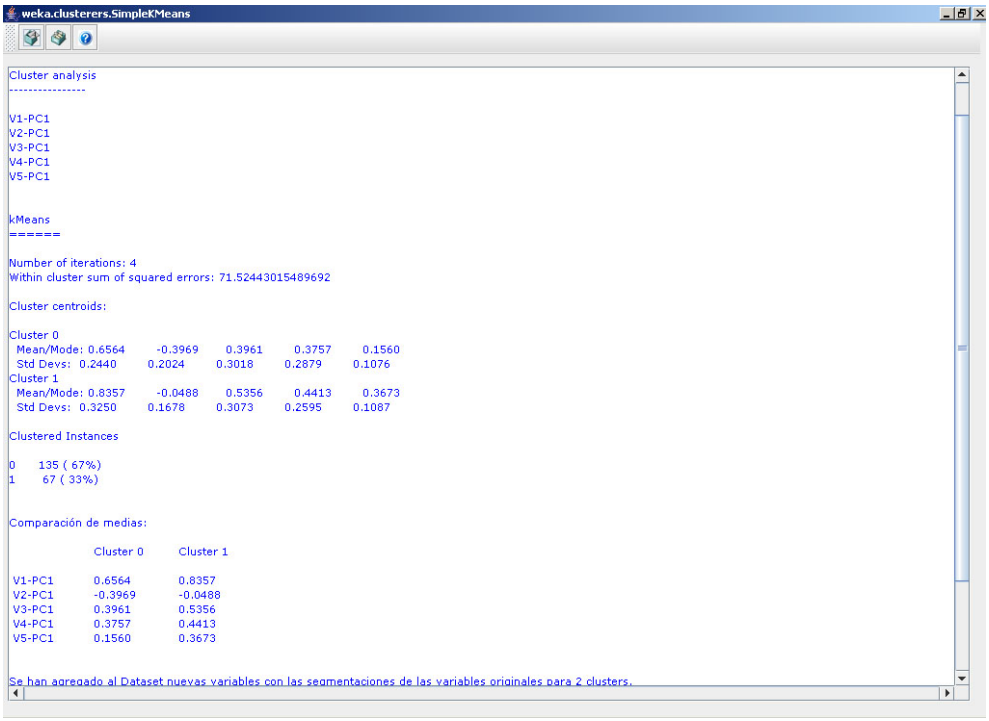
Figura *24. Ventana Análisis Cluster SimpleKMeans. Fuente Autores.*



1. En este campo se encuentran todas las variables disponibles cargadas en la sábana de datos, tanto las dependientes como las independientes.
2. Estos botones son los controles de desplazamiento de las variables hacia el campo dependiente, esto significa que en el campo **3** se pondrá la variable dependiente que se encuentra en la lista de variables, si es errado, se puede quitar la selección con botón que tiene el símbolo “<”.
3. Campo para ubicar la variable dependiente.
4. Estos botones son los controles de desplazamiento de las variables hacia el campo independiente, esto significa que en el campo **5** se pondrán las variables independientes que se encuentran en la lista de variables, si son errados, se puede quitar la selección con botón que tiene el símbolo “<”.
5. Campo para ubicar las variables independientes.
6. Este botón permite iniciar los análisis correspondientes al este proceso.
7. El *NumCluesters* indica la cantidad de clusters que el usuario desee que el prototipo proporcione.
8. *Seed* es la elección de la semilla para los valores aleatorios.

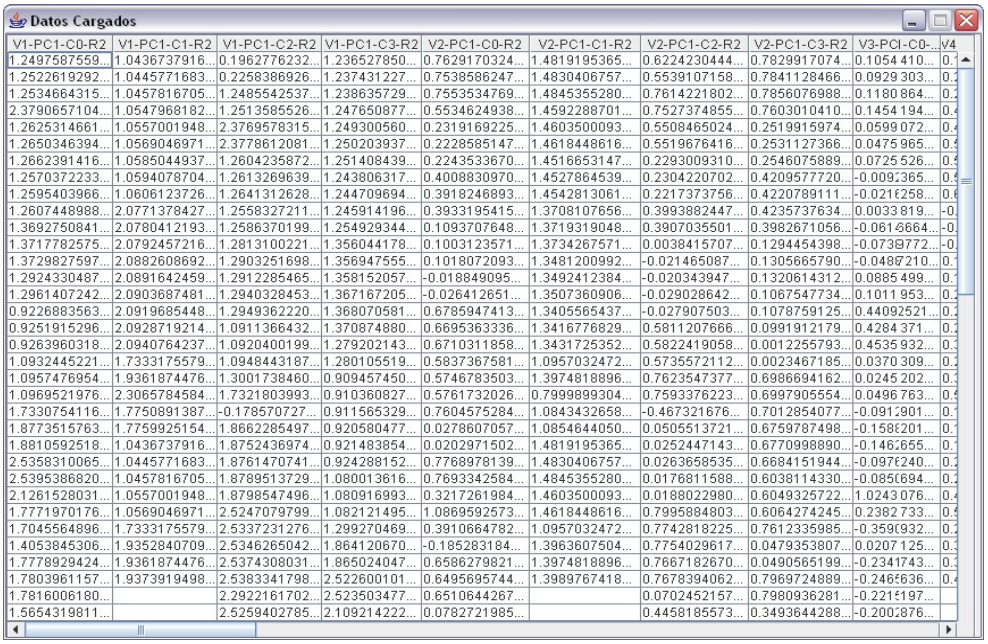
Finalmente al hacer clic sobre el botón analizar se obtiene la ventana de resultados de Cluster SimpleKMeans, Figura 25.

Figura *25. Ventana de Resultados Cluster SimpleKMeans. Fuente Autores.*



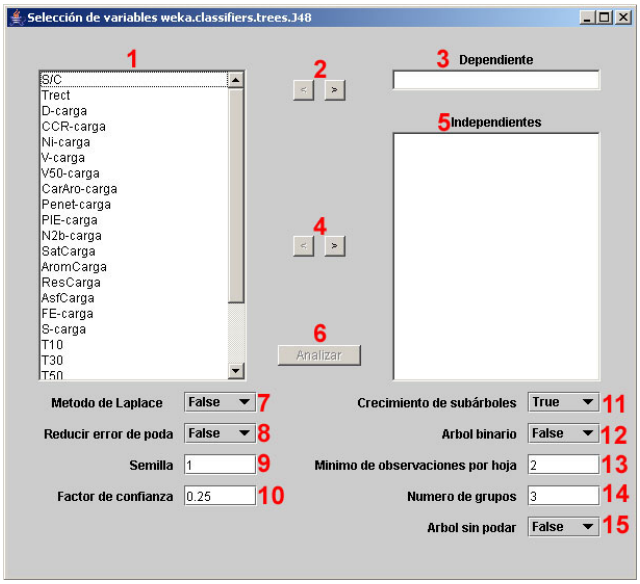
Similar a la técnica de componentes principales, la técnica de cluster genera nuevas variables, como se pude ver en la Figura 26.

Figura *26. Tabla con las nuevas variables generadas por Cluster. Fuente Autores.*



6. Árboles de decisión: Esta opción inicia el proceso de Árboles de decisión algoritmo J48, aplicado a los datos seleccionados previamente. Una vez ejecutado, seleccionando esta opción se puede ver la ventana de selección de variables para este proceso en la Figura 27.

Figura *27. Ventana Análisis Árboles J48. Fuente Autores.*

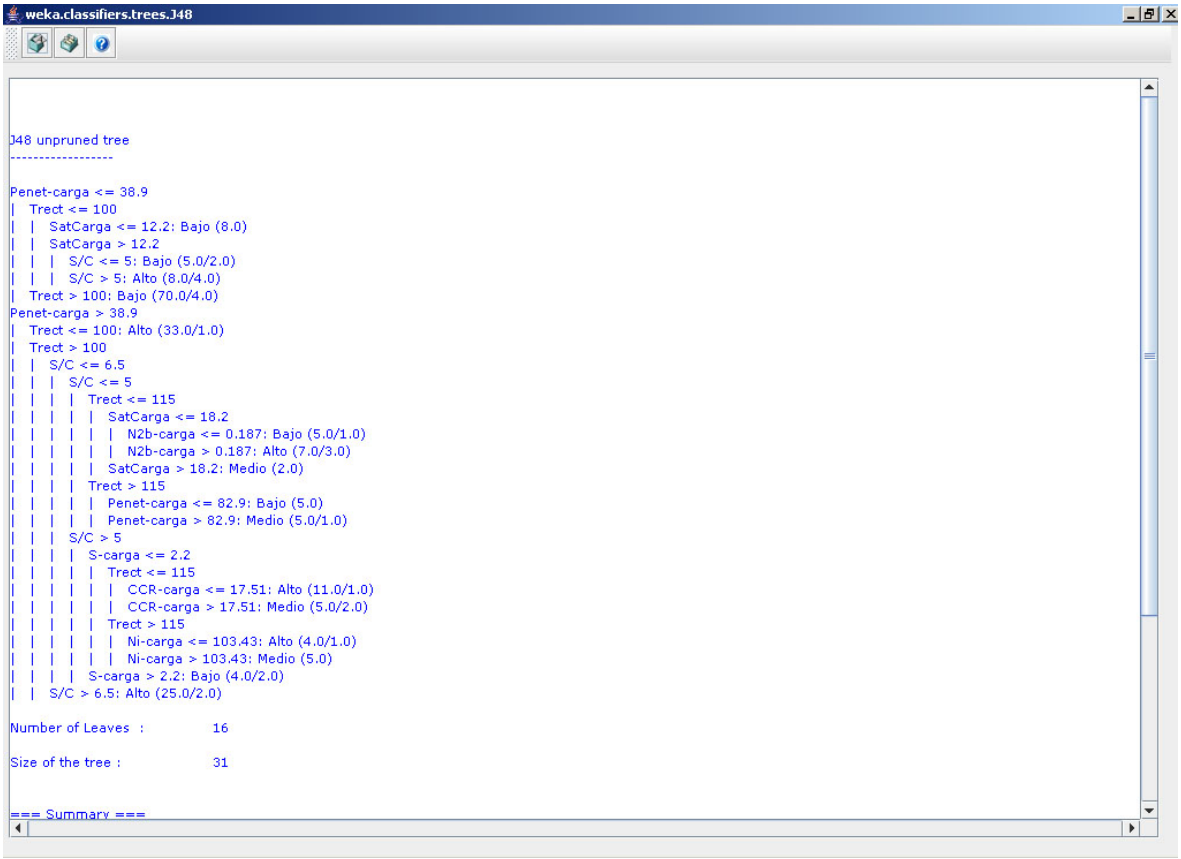


1. En este campo se encuentran todas las variables disponibles cargadas en la sábana de datos, tanto las dependientes como las independientes.
2. Estos botones son los controles de desplazamiento de las variables hacia el campo dependiente, esto significa que en el campo **3** se pondrá la variable dependiente que se encuentra en la lista de variables, si es errado, se puede quitar la selección con botón que tiene el símbolo “<”.
3. Campo para ubicar la variable dependiente.
4. Estos botones son los controles de desplazamiento de las variables hacia el campo independiente, esto significa que en el campo **5** se pondrán las variables independientes que se encuentran en la lista de variables, si son errados, se puede quitar la selección con botón que tiene el símbolo “<”.
5. Campo para ubicar las variables independientes.
6. Este botón permite iniciar los análisis correspondientes al este proceso.
7. El Método de *Laplace* se usa para mejorar la probabilidad estimada.
8. El campo Reducir el error de poda se usa para usar o no la opción descrita de los números de grupos.
9. *Seed* es la elección de la semilla para los valores aleatorios.
10. El campo Factor de Confianza es usado para estimar la poda de las ramas del árbol, cuales si, o cuales no.
11. Crecimiento de Subárboles es un campo que permite impedir el crecimiento de subárboles.

12. Árbol Binario es un campo con la opción para poder crear árboles binarios.
13. Mínimo de observaciones por hoja, indica el mínimo de casos por hoja.
14. Número de grupos utilizados en la reducción del error por el que las instancias son divididas en grupos para aplicar el clasificador a los grupos y así comparar.
15. Árbol sin podar**,** lo entrega con o sin podar.

Una vez ejecutado este análisis se puede referir a la ventana de resultados correspondiente a la Figura 28.

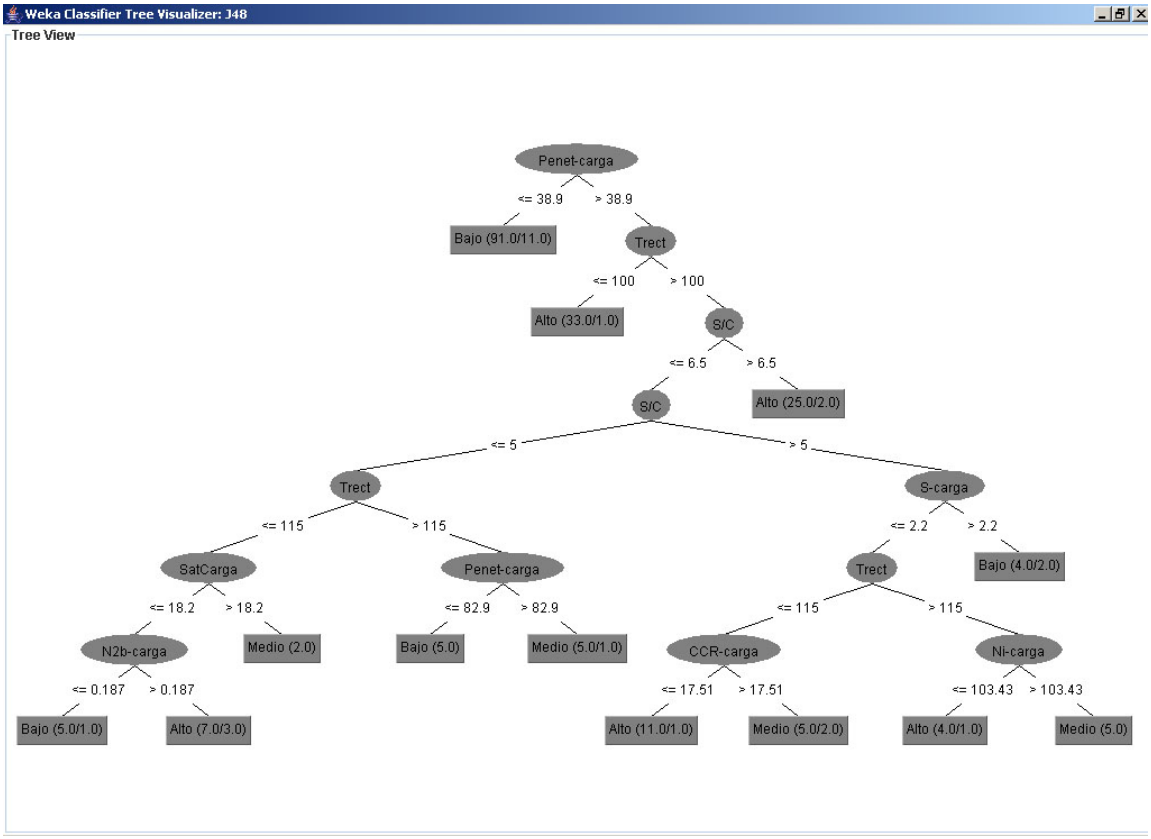
Figura *28. Ventana de Resultados de Árboles J48. Fuente Autor.*



Este prototipo cuenta con visualización de árboles en gráfico para su fácil interpretación como se aprecia en la Figura 29.

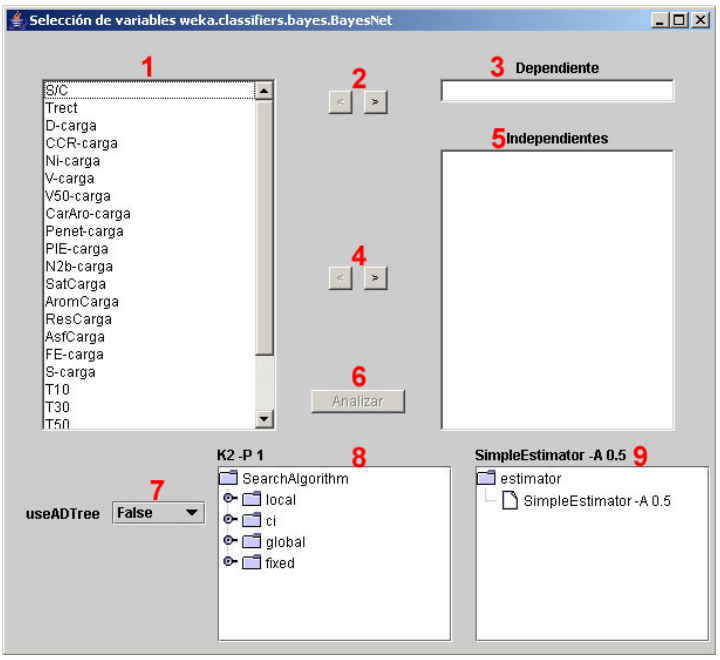
Para ampliar el árbol de debe maximizar la ventana, clic botón derecho en el espacio en blanco y seleccionar la opción ***fit to screen***.

Figura *29. Ventana de Resultados de Árboles J48 modelo gráfico. Fuente Autores.*



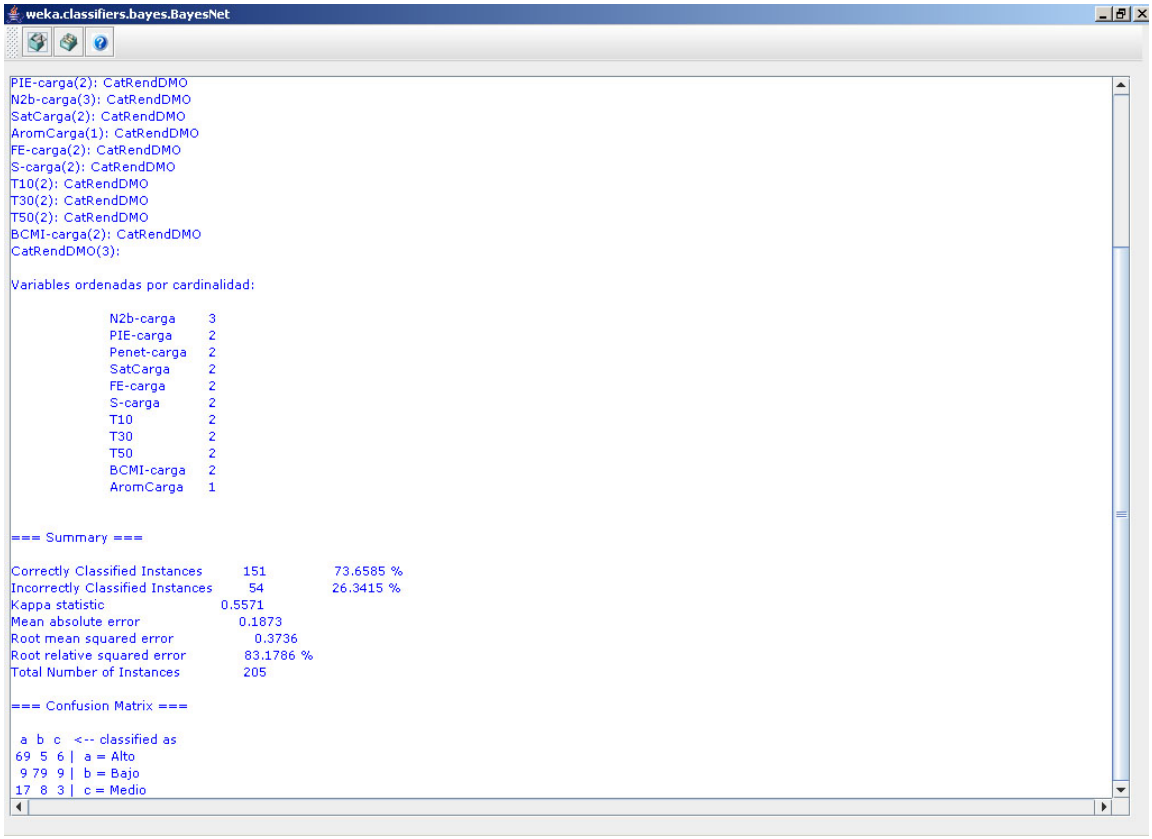
7. Redes Bayesianas: Esta opción inicia el proceso de redes bayesianas BayesNet, aplicado a los datos seleccionados previamente. Una vez ejecutado, seleccionando esta opción se puede ver la ventana de selección de variables para este proceso en la Figura 30.

Figura *30. Ventana Análisis Redes Bayesianas BayesNet. Fuente Autores.*



1. En este campo se encuentran todas las variables disponibles cargadas en la sábana de datos, tanto las dependientes como las independientes.
2. Estos botones son los controles de desplazamiento de las variables hacia el campo dependiente, esto significa que en el campo **3 s**e pondrá la variable dependiente que se encuentra en la lista de variables, si es errado, se puede quitar la selección con botón que tiene el símbolo “<”.
3. Campo para ubicar la variable dependiente.
4. Estos botones son los controles de desplazamiento de las variables hacia el campo independiente, esto significa que en el campo **5** se pondrán las variables independientes que se encuentran en la lista de variables, si son errados, se puede quitar la selección con botón que tiene el símbolo “<”.
5. Campo para ubicar las variables independientes.
6. Este botón permite iniciar los análisis correspondientes al este proceso.
7. *useADTree* se usa para saber si se hace o no una especie de árbol para la clasificación.
8. *SearchAlgorithm* Busca las relaciones.
9. *EstimatorAlgorithm* Calcula las relaciones.
Una vez ejecutado este análisis se puede referir a la ventana de resultados correspondiente a la Figura 31.

Figura *31. Ventana de Resultados de Redes Bayesianas BayesNet. Fuente Autores.*



8. GAMS (*General Algebraic Modeling System*): Esta opción inicia el proceso de análisis para sistemas no lineales, aplicado a los datos seleccionados previamente. Es importante resaltar que en la ruta se utiliza doble *slash* o *slash* invertido ya que es para una aplicación java, Figura 32.

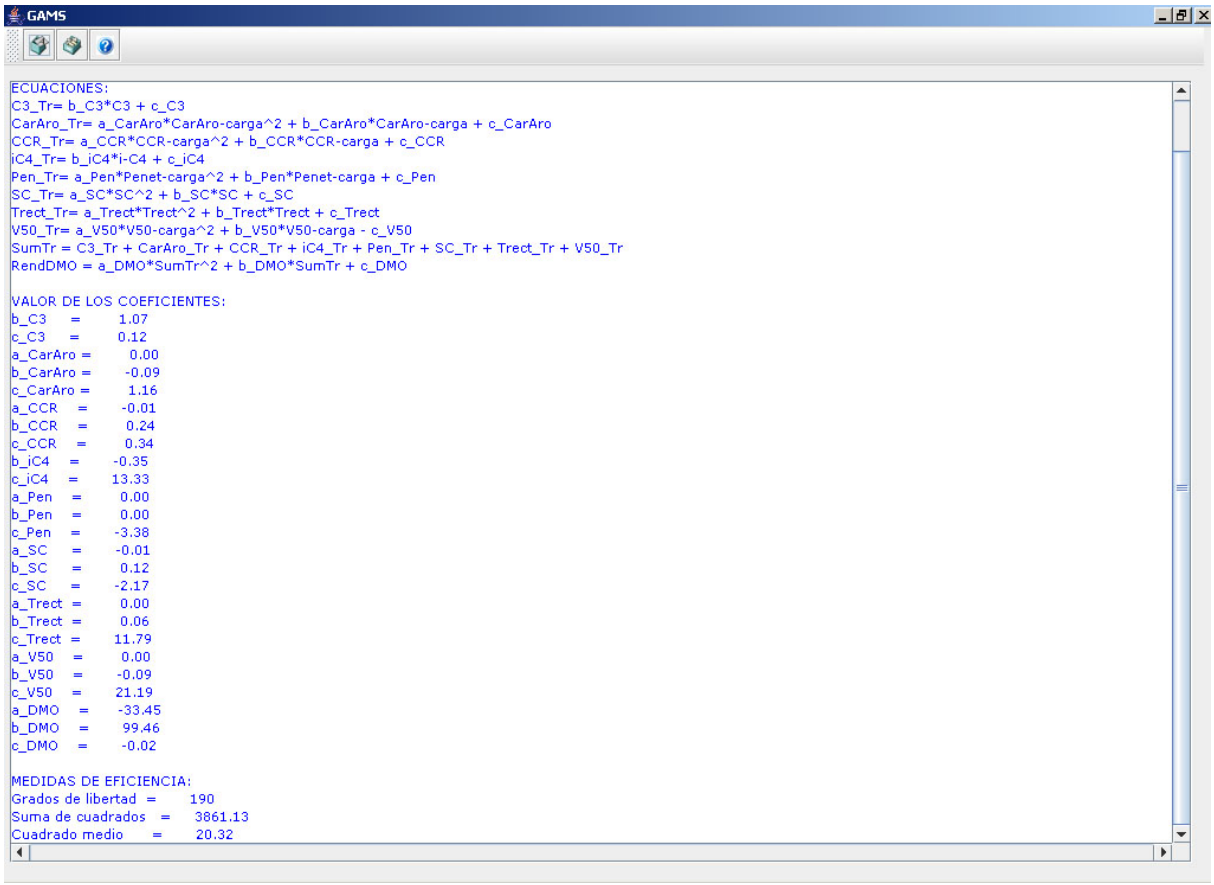
Figura *32 Formato del archivo GAMS.txt. Fuente Autores.*

```
C:\\Archivos de programa\\GAMS21.5\\gams.exe
C:\\Archivos de programa \\SPP\\solver\\DMO.gms
C:\\Archivos de programa \\SPP\\solver
```

GAMS genera un archivo de extensión .GMS con las variables correspondientes cargadas y hace un llamado a la aplicación para que procese estos datos. El modelo no lineal en una de sus líneas contiene un llamado o sentencia import que haga uso del archivo de datos de GAMS con la extensión GMS mencionada anteriormente y que se encuentra en la carpeta C:\\Archivos de programa \\SPP\\solver para a continuación mostrar los resultados del proceso al usuario. Es

vital para la aplicación que los modelos se encuentren en la carpeta mocionada, y los resultados serán almacenados en el archivo modelo.dat.

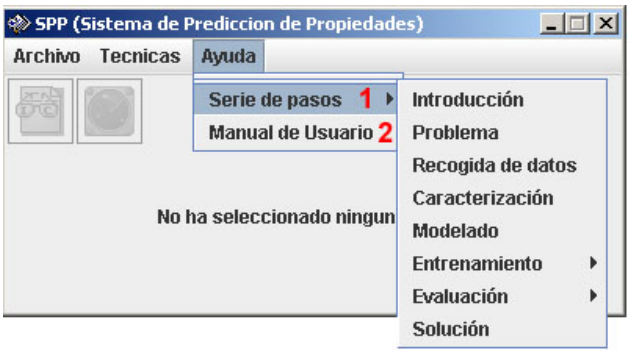
Figura *33: Resultados de la técnica GAMS. Fuente Autores.*



Menú Ayuda

Este menú se ha incluido en esta versión del prototipo para hacer más fácil el uso del mismo y no desamparar al usuario en sus dudas, la vista de la venta principal del menú ayuda, para desplegar este documento guía se puede ver en la grafica 34 con el numeral **2**.

Figura *34 Menú de ayuda. Fuente Autor.*



1. También se ha agregado una de las características más fuertes de esta aplicación al menú de ayuda, se trata del menú metodología, el cual proporciona un asistente de navegación que recomendara al usuario sobre la jerarquía y el orden de la serie de pasos que debe seguir para hacer una buena y sana practica de minera de datos, asistiéndolo todo el tiempo, y reduciendo así el error entre procesos, en un método que debe ser totalmente asistido.

Anexo H. Manual técnico

REQUERIMIENTOS MÍNIMOS DEL SISTEMA

Requerimientos mínimos de Hardware (para el SPP):

1. RAM 512 MB
2. Disco Duro 2 GB
3. Tarjeta de Red 10/100 Mbps Ethernet
4. Procesador Pentium IV

Requerimientos mínimos de software (para el SPP 2.0):

1. Sistema Operativo Windows XP ó superior
2. Máquina Virtual Java
3. Acceso a la base de datos SILAB

Requerimientos mínimos de Hardware (para GAMS):

1. RAM 256 MB
2. Disco Duro 100 MB
3. Procesador Pentium IV

Requerimientos mínimos de Software (para GAMS):

1. Sistema Operativo Windows 98 o superior

Requerimientos mínimos de Hardware (para *WEKA*):

1. RAM 256 MB
2. Disco Duro 60 MB
3. Procesador Pentium IV

¿CÓMO INSTALAR EL PROTOTIPO?

Antes de iniciar cualquier procedimiento de instalación es necesario que el usuario haya instalado con anterioridad el software gams, el cual debe estar instalado en la ruta C:/Archivos de programa

Si no se encuentra en esa ruta, una vez instale el prototipo deberá actualizar las rutas en la carpeta GAMS el archivo config que se encuentra en la siguiente ruta C:/Archivos de programa/SPP/GAMS/config

El archivo config consta de la configuración para el análisis de la técnica GAMS.

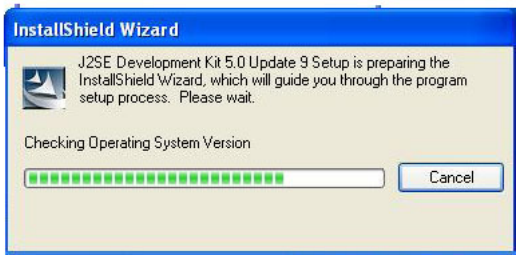
```
C:\\Archivos de programa\\GAMS21.5\\gams.exe
C:\\Archivos de programa\\SPP\\solver\\DMO.gms
C:\\Archivos de programa\\SPP\\solver
```

- La primera línea indica la ruta al programa GAMS.
- Las otras dos líneas no deben ser modificadas. Si desea cambiar el modelo debe sobrescribir el archivo DMO.gms

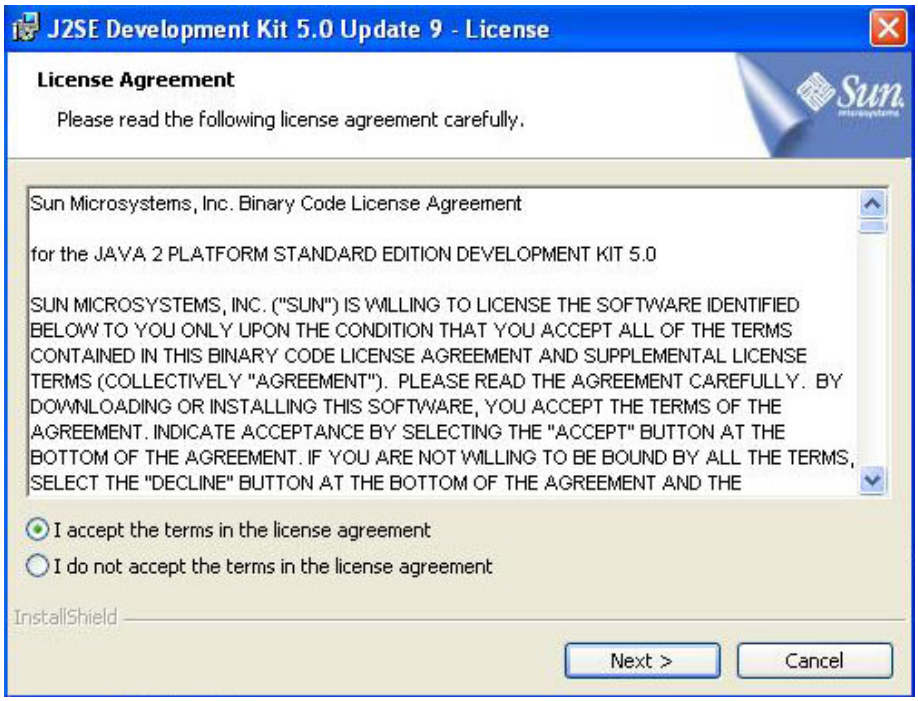
Cuando se abre el CD, automáticamente se despliega la ventana principal para instalar los componentes necesarios para que el prototipo funcione de manera correcta. Esta ventana puede verse en la Figura 1.

Figura 1. Ventana principal de instalación

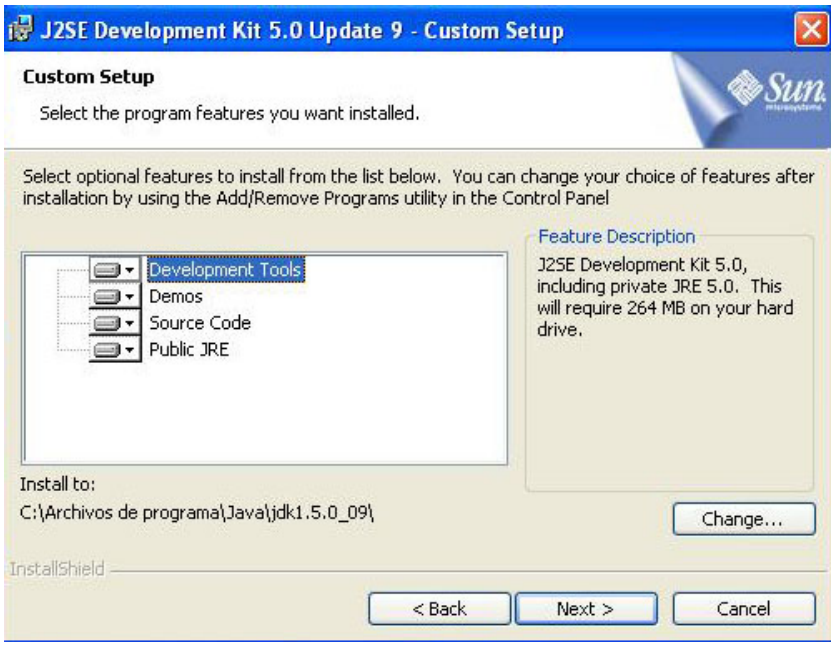
Se recomienda instalar primero la máquina virtual de java, es decir, haciendo clic en el icono J2DK en la ventana principal. Esto se hace de la siguiente manera:



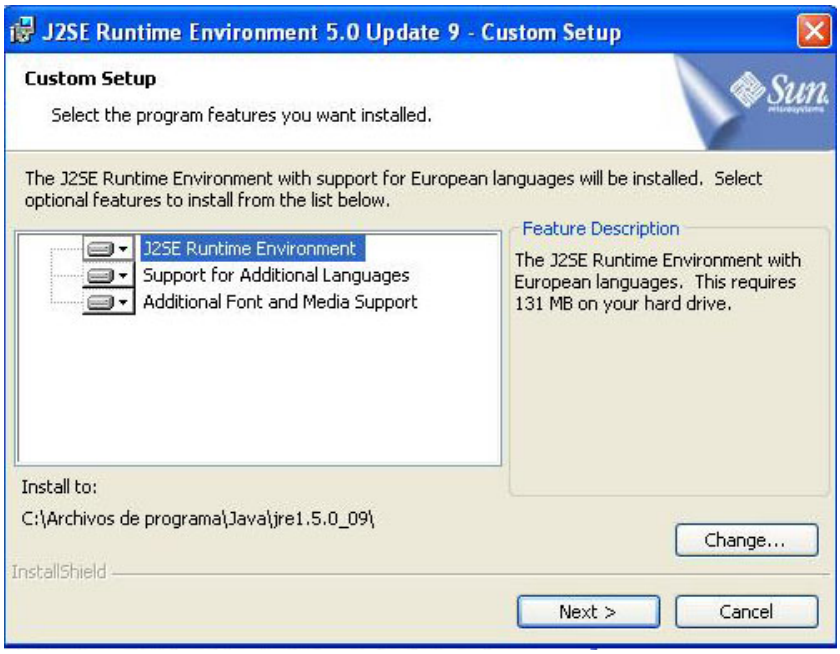
Es la primera ventana que se encuentra cuando se activa la opción J2DK, después se continúa con la aceptación de la licencia para el uso de la máquina virtual por parte de la empresa desarrolladora Sun Microsystems.



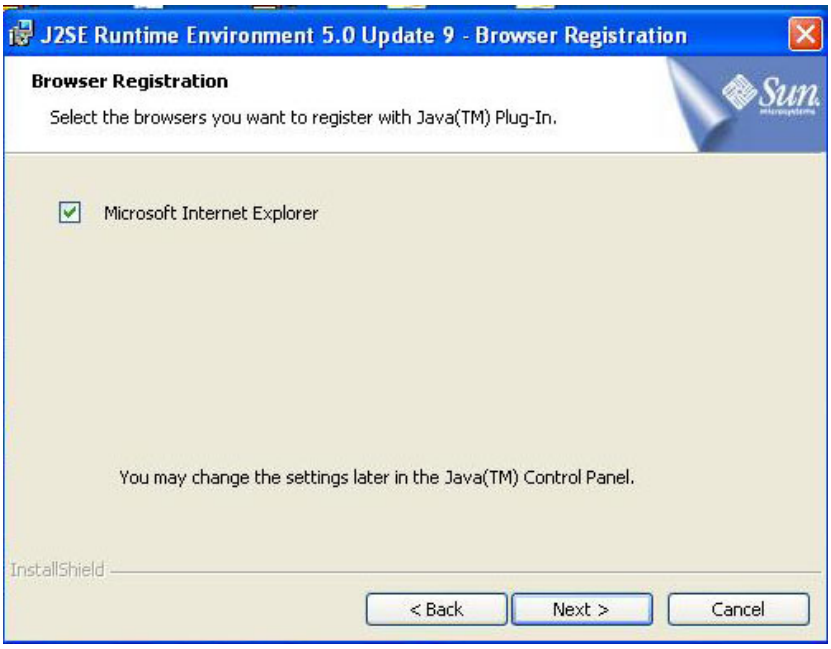
En esta opción el usuario debe seleccionar el ítem "*I accept the terms in the license agreement*" y dar clic en el botón "*next*". Para escoger las características a instalar.



El usuario solo debe dar clic en el botón "next". La interfaz inicia el proceso de instalación la cual demora aproximadamente 5 minutos o más dependiendo del ordenador. Una vez termina, se continúa con la siguiente ventana.



En esta opción se instalara el J2SE el ambiente de ejecución, el usuario debe dar debe dar clic en el botón “next”. Cuando instale estos componentes avisa al usuario que usará Internet Explorer como el *browser* por defecto. Si el usuario usa más de un *browser* estos saldrán y se debe escoger el de más agrado.

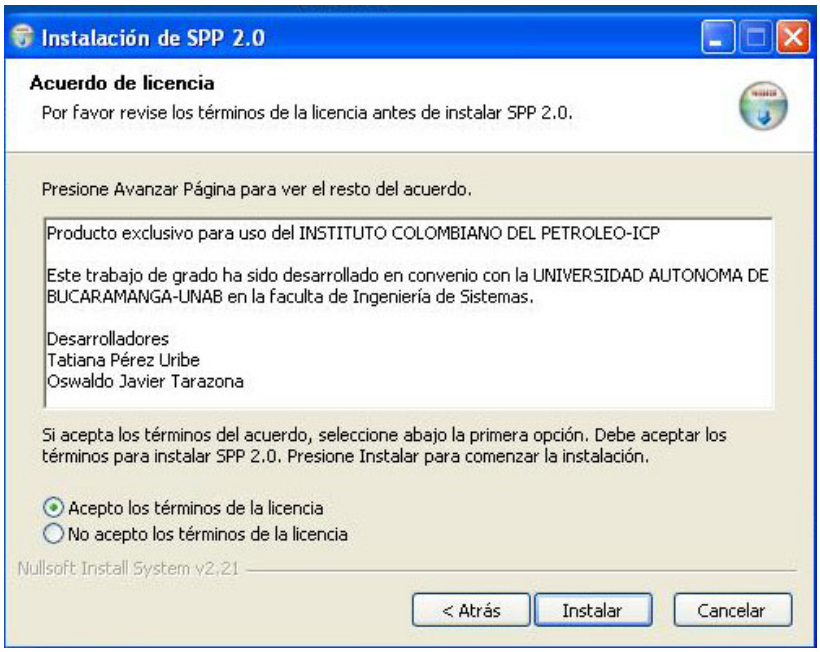


Se debe dar clic en el botón “*next*” y se da por terminada la instalación dando clic en el botón “*finish*”.

Ahora, se instalará el prototipo dando clic en SPP 2.0.

Y saldrá la siguiente ventana

Para iniciar la instalación el usuario debe dar clic en el botón “siguiente”. Para dar paso a la aceptación de la licencia.

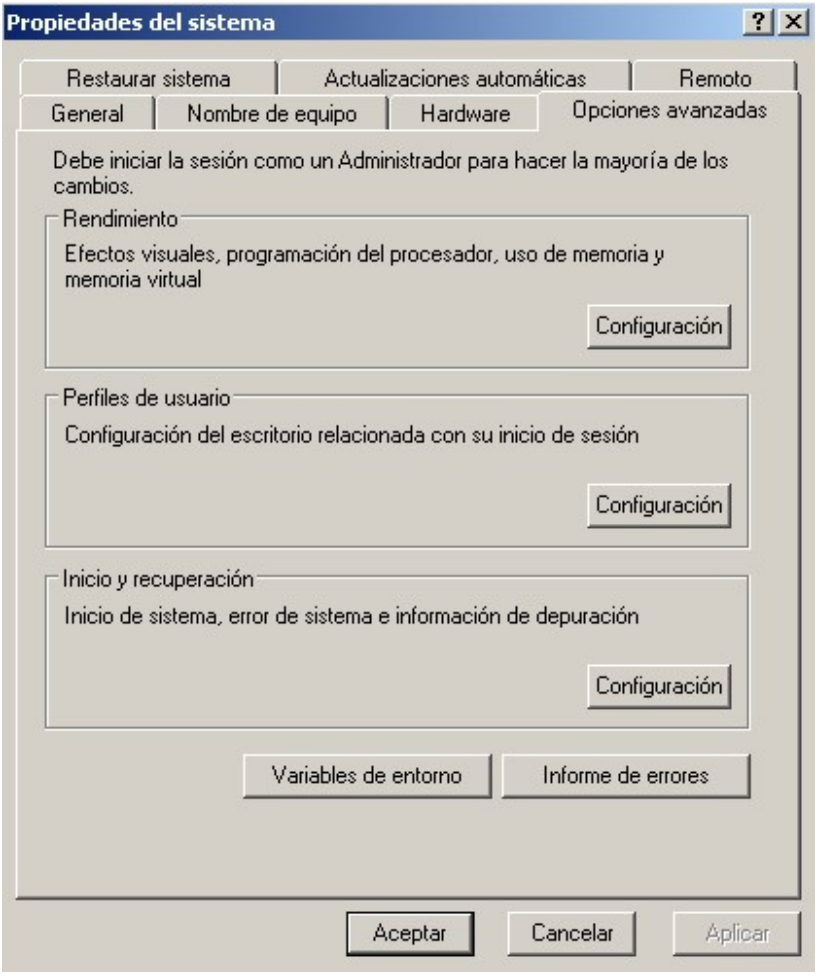


El usuario debe escoger “Acepto los términos de la licencia” y a continuación clic en el botón “instalar”.

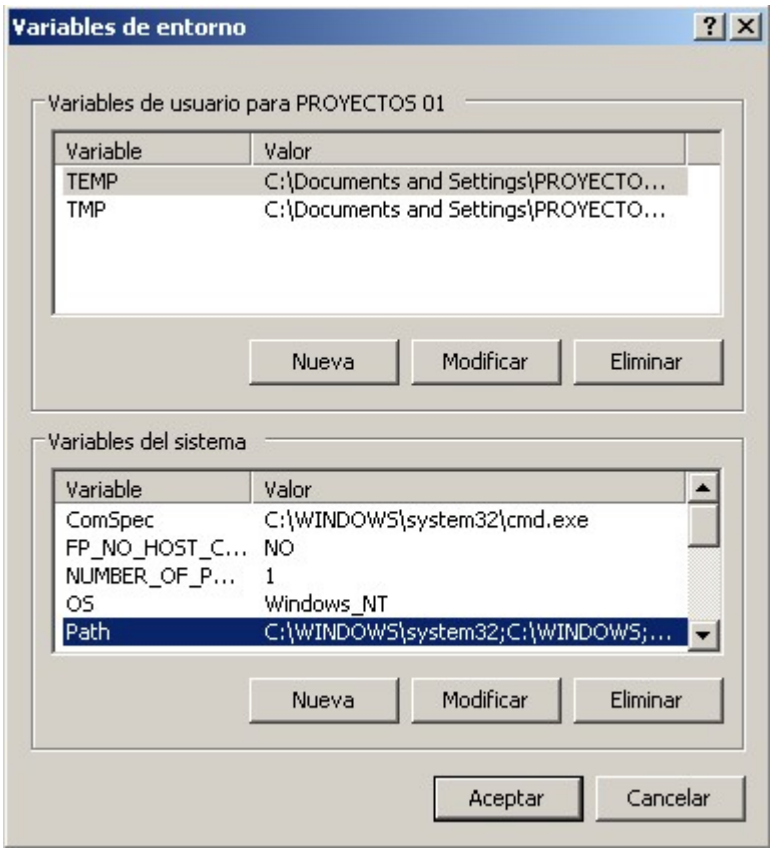
Se da por terminada la instalación cuando se de clic en el botón “Terminar”.

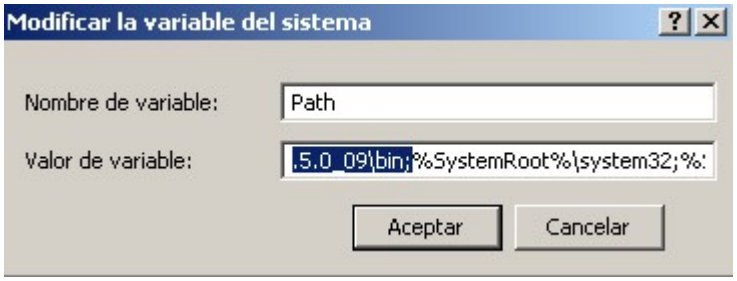
Cuando se intente correr la aplicación y está arroje la siguiente ventana

Se debe ir a **configuración/panel de control/sistema** opciones avanzadas buscar la pestaña o botón **variables de entorno.**



Buscar la variable de entorno path, dar clic en modificar y sin borrar, es decir, poner al principio la siguiente ruta **C:\Archivos de programa\Java\jdk1.5.0_09\bin;** seguida del **;**.


Variables de entorno
Variables de usuario para PROYECTOS 01
Variable
Valor
TEMP
C:\Documents and Settings\PROYECTO...
TMP
C:\Documents and Settings\PROYECTO...
Nueva
Modificar
Eliminar
Variables del sistema
Variable
Valor
ComSpec
C:\WINDOWS\system32\cmd.exe
FP_NO_HOST_C...
NO
NUMBER_OF_P...
1
OS
Windows_NT
Path
C:\WINDOWS\system32;C:\WINDOWS;...
Nueva
Modificar
Eliminar
Aceptar
Cancelar


Modificar la variable del sistema
Nombre de variable:
Path
Valor de variable:
.5.0_09\bin;%SystemRoot%\system32;%:
Aceptar
Cancelar