Universidad de Valparaíso
Facultad de Ingeniería

Escuela de Ingeniería Civil Industrial

Modelo para Optimización de Rutas de Transporte Secundario en Agrosuper Comercial

por

Daniela Paz Navarro Cifuentes

Julia Carolina Aracena Orsola

Trabajo de Título para optar al Grado de
Licenciado en Ciencias de la Ingeniería y título de
Ingeniero Civil Industrial

Profesor Guía: Boris Cariqueo

Junio, 2017

Agradecimientos

Gracias por el apoyo incondicional de siempre: a mi familia y amigos, son quienes mas amo en este mundo.

DDP

Agradezco a mi familia y amigos quienes siempre me apoyaron y, no puedo dejar fuera a mi compañera y amiga Daniela, quien hizo posible la existencia de esta instancia.

Julia Aracena

Dedicatorias

Dedico este trabajo para Todos quienes siempre queremos más, el esfuerzo vale la pena aunque tarde en llegar.

DDP

Dedico este trabajo a todos quienes siempre me apoyaron incondicionalmente e insistieron en que esto era posible, a mis familiares y amigos.

Julia Aracena

Índice

Glosario

Child Care: disponibilidad de dejar los niños a cuidado

Ease: facilidad del centro para concertar cita y eficiencia de la misma

Health: salud del paciente

KDD: Knowledge Discovery from Databases

Need: convicción del paciente que la visita es importante

No-Show: indica si el paciente no se ha pasado por el médico durante el último

Sick Time: si el paciente está trabajando, puede darse de baja

Satisfaction: satisfacción del cliente con su médico

Transportation: disponibilidad de transporte del paciente al centro

/

Listado de abreviaturas y siglas

Add: Añadir

ASCII: código estándar estadounidense para el intercambio de información (American Standard code for Information Interchange)

CN: cliente nuevo

FDV: Fuerza de venta

ICR: Índice de complejidad de ruta

JDBC: permite la ejecución de operaciones sobre datos desde el lenguaje de programación Java (Java Database Connectivity)

JF: Jefe de frigorífico sucursal

JO: Jefe de operaciones

Req.SS: requerimiento de servicio

SS Cliente: Servicio cliente

Listas de figuras

Lista de tablas

Resumen

Esta tesis se basa en la empresa comercializadora de alimentos Agrosuper, con la finalidad de determinar una forma para que el transportista que trabaja como distribuidor, sea capaz de vender la mercadería extra que lleva en su vehículo, optimizando su tiempo y ruta junto a ser atractivo el hecho de querer vender más.

El despacho y distribución de los distintos productos (aves, cecinas, cerdos, elaborados, hortalizas y frutas, pavos, salmón) se realiza con transportistas externos, quienes prestan servicio para abastecer a la cartera de clientes existente. Los tipos de clientes atendidos por Agrosuper se clasifican en supermercados, industriales, food services, grandes clientes y canal tradicional.

El modelo de comisión de los transportistas está conformada por un polinomio que considera los siguientes ítems: costo fijo, corresponde al cumplimiento de los requerimientos básicos del camión solicitados por la empresa y los organismos gubernamentales; costo variable corresponde a la cantidad de kilómetros por ruta más el desgaste asociado del vehículo (mantenciones); Índice de complejidad de la ruta, relacionado con la cantidad de clientes atendidos mes, a mayor cantidad de clientes entregado, mayor comisión; y gestión del transportista la cual mide el nivel de servicio, el cual mide que la cantidad de producto despachado a los clientes sea igual a lo facturado, la devolución de producto y el precio promedio de venta del mismo.

Actualmente se ha detectado que este modelo no es atractivo para los transportistas, ya que no incentiva a transportar producto independiente de la cantidad de clientes a atender y además, trabaja con rutas fijas, por lo que se trabaja en esta tesis en determinar un modelo de comisiones que permita incentivar las rutas variables sin aumentar el neto total de las comisiones pagadas. Para ello se utiliza el software llamado *WEKA* y *STATA,* los cuales son capaces de trabajar con bases de datos y entregar una combinación de modelos, datos en múltiples gráficas, además de dar la posibilidad de realizar pruebas estadísticas que ayudarán a determinar el mejor polinomio para el vendedor y la empresa.

Luego de trabajar con ambos programas, se obtuvo dos rutas de distribución. La primera ruta (llamada A) se determinó un polinomio para entrega exclusiva a supermercados y una segunda ruta (llamada B) para entrega a todo tipo de clientes. Ambas rutas consideran

dentro de su polinomio costos fijos, transacciones, kilómetros por mes y en el caso de la segunda ruta se agrega el índice de complejidad de ruta.

Introducción

1.1. Antecedentes de la empresa

Agrosuper es un holding de empresas alimentarias chilenas, dedicada principalmente a la producción, distribución y comercialización de alimentos frescos y congelados de cerdo, aves (pollos y pavos), salmones, productos procesados (cecinas), elaborados (hamburguesas) y hortalizas congeladas.

La historia de Agrosuper se remonta al año 1955 con la producción de huevos en la localidad de Doñihue, VI Región. Cinco años más tarde, Gonzalo Vial, fundador de la compañía, decide expandir el negocio hacia la producción y comercialización de animales vivos.

En el año 1974, Agrosuper amplía su negocio al procesamiento y comercialización de carne de pollo, lo que marca el inicio de las actividades que desarrolla actualmente a través de la marca Super Pollo.

En el año 1983 Agrosuper ingresa al negocio del cerdo aprovechando su experiencia en la crianza de animales vivos y la infraestructura disponible. A los pocos años, la Compañía amplía el negocio al procesamiento y comercialización de carne de cerdo a través de la marca Super Cerdo.

Durante el año 1989, la compañía ingresa al negocio de la elaboración de cecinas pensando en dar un mayor valor agregado a la carne de pollo y cerdo y aprovechar las posibles sinergias en distribución y comercialización. En este mismo año, se comienza con la producción y comercialización de truchas y salmón atlántico, instalándose como pionera en la zona del canal Puyuhuapi, XI Región.

A partir del año 1990, Agrosuper inicia su proceso de expansión internacional vía la exportación de productos provenientes de los negocios de pollos, cerdos y salmones.

En el año 1996, Agrosuper ingresa a la propiedad de Sopraval con el fin de aprovechar toda su experiencia en la crianza de animales en el negocio de pavos y consolidar potenciales sinergias en la producción y comercialización de sus productos.

Debido a la experiencia en el negocio de pollos y con el objetivo de crecer en el mercado local, Agrosuper adquiere en el año 2000 Pollos King, lo cual le permite captar una mayor variedad de clientes.

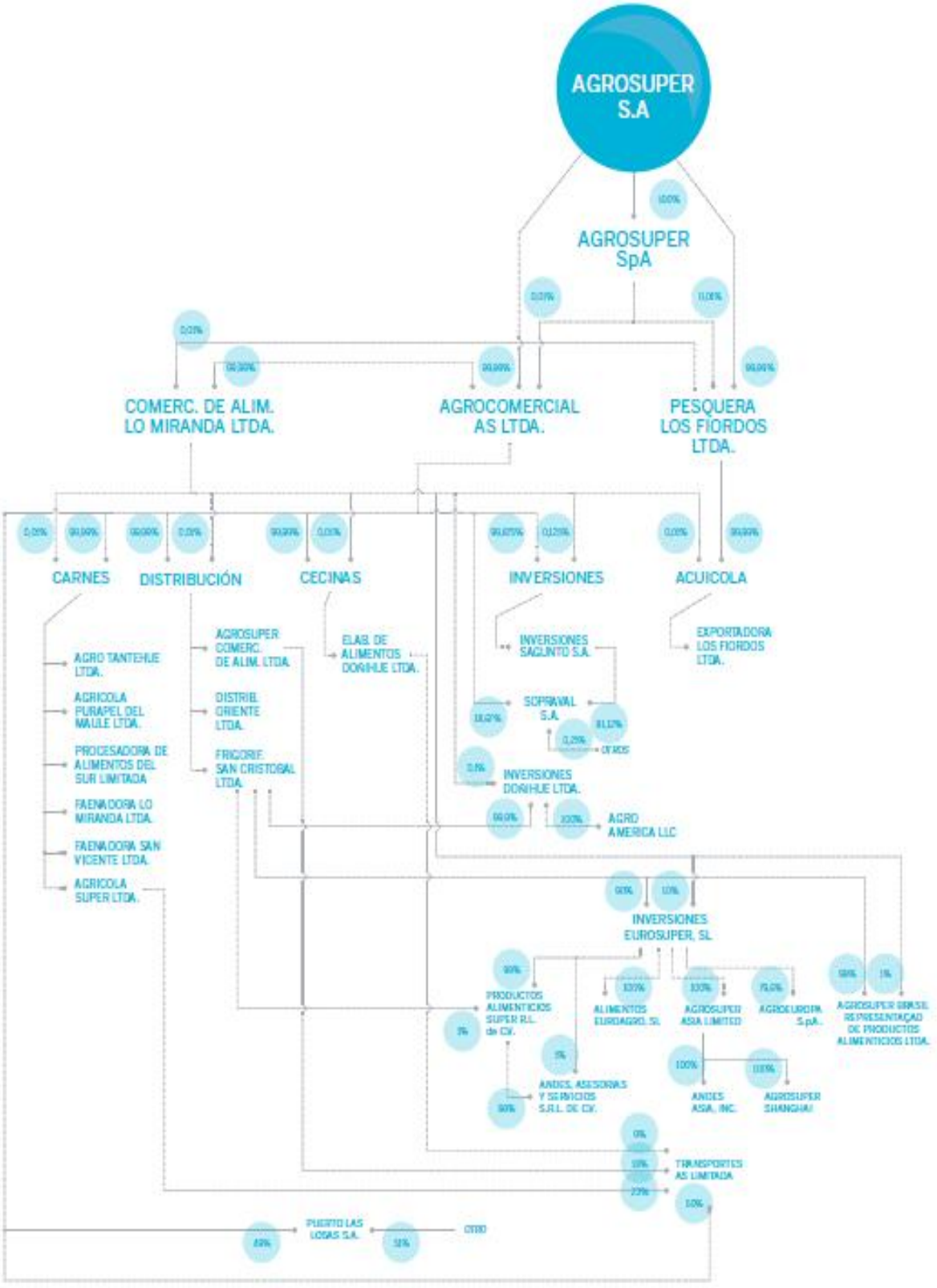
A partir del año 2002 comienza un proceso de apertura de oficinas comerciales propias en los principales mercados donde participaba la compañía, con el fin de entregar una atención personalizada a sus clientes y crear alianzas con los distribuidores locales. Inicialmente se instala en Italia, luego en el 2003 en Estados Unidos y Reino Unido, el 2004 en Japón, el 2005 en México y China en el año 2009.

Como una forma de continuar con un crecimiento constante en su negocio de cerdos, en el año 2005 Agrosuper inicia el desarrollo del “Proyecto agroindustrial valle del huasco", el cual tiene como objetivo duplicar su producción de carne de cerdo.

Durante el año 2011, Agrosuper adquiere la participación accionaria de Sopraval, llegando así a un 99,7% de la propiedad.

En la actualidad Agrosuper es la principal empresa productora de proteína animal y alimentos frescos y congelados de Chile, con una destacada presencia en el mercado local y de exportación.

Figura 1: Esquema de la empresa Agrosuper



1.2 Visión, misión y pilares estratégicos

Agrosuper combina una fuerte vinculación con el campo, con un poderoso componente de eficiencia industrial de estándares internacionales, los que permiten competir en los mercados nacionales y extranjeros. Durante sus 60 años de vida, ha experimentado un crecimiento sostenido a través del crecimiento orgánico de las operaciones, aplicando en todos los procesos, los más altos estándares productivos y de calidad.

Su trabajo está basado en ser una empresa líder con altos estándares de calidad pensando siempre en innovar y entregar lo mejor a sus clientes. Éste trabajo se ve claramente reflejado en su visión y misión.

La Visión de la empresa es: Ser una empresa líder a nivel mundial destacada por sus productos, buenas prácticas, innovación, trayectoria y excelencia en sus procesos. Caracterizada por la seriedad y sustentabilidad de su gestión, y deseada como uno de los mejores lugares para trabajar.

La Misión de la empresa es: Procurar alimentos para Chile y el mundo en forma sustentable e innovadora, creando valor junto a nuestros consumidores, trabajadores, inversionistas, vecinos y proveedores, bajo los más altos estándares de calidad, inocuidad y excelencia.

El crecimiento de Agrosuper en Chile y en el extranjero ha estado fundamentado en cuatro pilares:

- Ser líderes del desarrollo agroindustrial de la región.

Desde la región de O'Higgins, Agrosuper basa su liderazgo productivo y comercial en el desarrollo sustentable de sus productos y marcas junto a las comunidades vecinas, a través de un vínculo cercano y transparente.

- Ser productores de clase mundial.

Agrosuper cuenta con un modelo de integración vertical, lo que le permite mantener el control y trazabilidad de sus insumos y productos, maximizando las economías de escala y facilitando la diversificación de los alimentos que produce.

- Extensa red de distribución en Chile.

La extensa red de distribución y canales de venta que Agrosuper ha construido, le permite acceder a más del 98% de la población del país con una cartera diversificada de clientes, acercando las ventas al consumidor final.

- Replicar su exitoso modelo de posicionamiento como motor de crecimiento.

Las exportaciones se han fortalecido a lo largo del tiempo, posicionándose como una fuente de crecimiento importante para la empresa. Asegurar este nivel de crecimiento requiere tener un estricto control sobre las medidas sanitarias del proceso productivo y de comercialización, con el objeto de minimizar el riesgo de contagio y conservar la relación con los mercados más exigentes del mundo.

Figura 2: Hitos 2015 de la empresa

2. Descripción modelo de distribución

La empresa Agrosuper Comercializadora de Alimentos cuenta con 29 sucursales a lo largo del país desde Arica hasta Punta Arenas, las cuales son abastecidas (distribución primaria) desde las tres plantas de producción de cerdo y ave ubicadas en la localidad de Doñihue, VI Región; una planta de pavo ubicada en La Calera, V Región, y una planta de procesados (elaborados, hortalizas, salmón y cecinas) ubicada en San Pablo, Región Metropolitana.

Este producto se despacha desde las sucursales a los diferentes tipos de clientes, los cuales son:

- Supermercados
- Industriales
- Food services
- Grandes clientes
- Canal tradicional

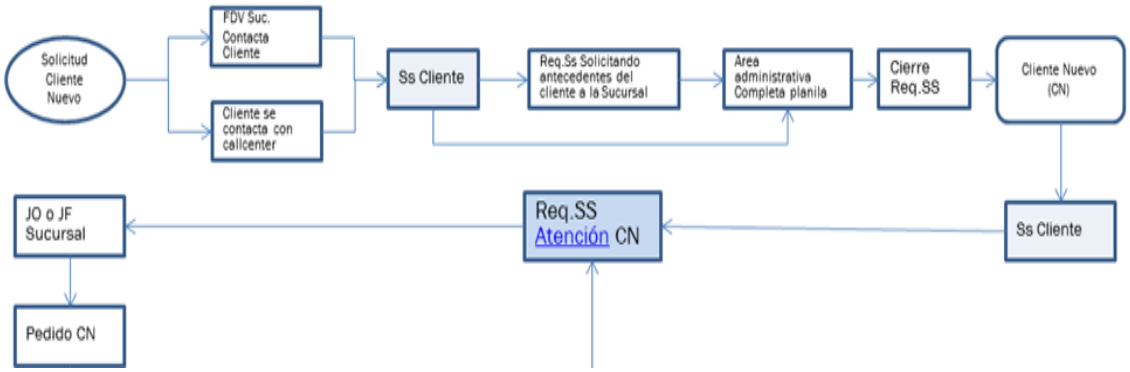
El despacho se realiza por transportistas externos, los que prestan servicio a la sucursal. En la actualidad existen en Agrosuper 392 camiones, correspondientes a 232 empresas transportistas.

2.1 Diagrama distribución secundaria

Las rutas en las sucursales son fijas, es decir, cada camión cuenta con una “cartera de clientes” distribuido en una zona específica, designada de manera aleatoria ya que no cuentan con un mapeo de los clientes por zona. Estos clientes pueden ser atendidos de forma diaria (lunes a sábado) independiente de la cantidad que hayan solicitado, porque no existe un pedido mínimo.

La creación de nuevos clientes se realiza de la siguiente manera:

Figura 3: Flujo creación de clientes



FDV: Fuerza de Venta; **Ss Cliente:** Servicio Cliente; **Req.SS:** requerimiento de servicio; **CN:** cliente nuevo; **JO – JF:** Jefe de Operaciones – Jefe de Frigorífico Sucursal

Cuando el cliente es creado por las vías anteriormente mencionadas, es el JO o JF de la sucursal el encargado de designar este cliente a algún transportista utilizando los siguientes criterios:

- Zona de reparto del transportista: cada transportista tiene una zona determinada (comunas) en donde reparte los pedidos, por lo tanto la asignación se realiza sólo en base a la dirección del nuevo cliente (ubicación geográfica).
- Capacidad máxima de carga de camión: los camiones tienen una capacidad de carga diaria de 8 pallets, equivalente a más menos 6.400 kilos día (mix de productos), es decir, 150 toneladas mes, por lo que si esta capacidad es excedida el nuevo cliente debe ser asignado a otro transportista que haga alguna ruta similar.
- Frecuencia de entrega de clientes supermercado y ventanas horarias: este tipo de cliente cuenta con una frecuencia de entrega en base a los requerimientos de las distintas cadenas (CCS, Walmart SMU y cadenas regionales), capacidades de almacenamiento de producto (cámaras) y comportamientos de ventas de los clientes (estacionalidades), por lo que restringe el horario de entrega de los transportistas y la cantidad de clientes a atender, ya que estos pedidos son los de mayor tonelajes y volúmenes.
- Camiones de reparto “exclusivos”: para supermercado, existen camiones exclusivamente designados a atender este segmento, por lo que no se puede asignar otro tipo de clientes.

- Canal tradicional, food service, industriales y grandes clientes: estos canales no cuentan con una frecuencia establecida, por lo que existe un número máximo de atender diariamente de alrededor 50 clientes

2.2 Sistema de comisión transportistas

El modelo de pago de los transportistas está conformado por los siguientes ítems:

a) Tarifas por sector y tipo de cliente:

Tabla 1: Valores en pesos a paga por sector y tipo de cliente

Sector	
Aves	$ 9
Cecinas	$ 8
Cerdos	$ 8
Elaborados	$ 10
Hortalizas y Frutas	$ 10
Pavos	$ 8
Salmón	$ 12

Tipo Cliente	
Consumidor	$ 5
Foodservice	$ 3
Grandes Clientes	$ 1
Industriales	$ 1
Supemercados	$ 2
Tradicional	$ 5

Estos valores son los basales para el cálculo del modelo de comisiones, y la forma en que se realiza es la siguiente:

Tabla 2: Datos para ejemplo de cálculo de comisión

Vendedor	Sector	Kilos Mes	Tipo Cliente	$ Sector	$ Cliente
V0001	Pollo	10.000	Supermercado	9	2

Por lo tanto: **10.000 * (9+2) = $110.000**

b) Costos fijos: Este ítem corresponde a todos los requerimientos básicos que necesita un camión para poder entrar a trabajar a Agrosuper, los cuáles son:

Tabla 3: Ítems considerados en costos fijos

Ítem
Permiso circulación
Seguro obligatorio
Certificado SAG
Revisión técnica
Uniforme
Teléfono
Colación
Sueldo peoneta
Mantención equipo de frío
Depreciación camión
Seguro camión
Equipo de GPS
Contador auditor
Provisión vacaciones
Provisión desahucio
Sueldo chofer
Imposición chofer
Imposición peoneta

Estos costos son pagados en un 85% del monto total y varía por región, producto de la diferenciación de sueldos de chofer y peoneta a nivel nacional. Este ítem se paga mensualmente, independiente de las ventas, por lo que pasa a ser el sueldo base del transportista.

Este indicador pesa un 49,5% con respecto a la comisión total.

c) Costos variables: Está asociado al desgaste del camión producto de los kilómetros recorridos en la ruta, lo cual está definido de la siguiente manera:

Tabla 4: Modo de pago de costo variable

Diesel bruto	**$ 480**

Item	**Cantidad**	**Valor**	**Cada cuantos Km**	**Valor por Km**
Diesel	1	$ 403	4	$ 101
Neumáticos	6	$ 130.000	45.000	$ 17
Cambio aceite y 3 filtros	1	$ 70.000	5.000	$ 14
Mantención preventiva	1	$ 160.000	10.000	$ 16

Por KM	**$ 148**

El valor pagado de pesos por kilómetros recorridos, varía mensualmente ya que se analiza en base al valor del precio del petróleo real.

Los kilómetros recorridos mes, son obtenidos desde el equipo GPS con que cuenta cada camión idividualmente.

Este indicador pesa un 12% con respecto a la comisión total.

d) Participación: Está asociado al aporte en kilos entregados por cada transportista con respecto a la venta total de la sucursal.

Tabla 5: Ejemplo participación transportistas sucursal Arica

Nom_Centro	**Cod_Vendedor**	**Kilos_Venta**	**Participación**	**$_Participación**
Arica	V1021	60.636	13,9%	$ 74.017
	V1740	63.708	14,6%	$ 82.666
	V1887	84.307	19,3%	$ 143.111
	V2118	70.743	16,2%	$ 102.919
	V2355	93.827	21,5%	$ 176.255
	V2356	63.159	14,5%	$ 80.878
Total Arica		**436.380**	**100%**	**$ 659.846**

La forma de pago se realiza mediante el cálculo de los kilos entregados mes por sector y tipo de cliente, de la base de tabla de tarifas expuestas anteriormente, por el porcentaje de participación:

Kilos entregados * (sector + tipo de cliente) * % participación

Este indicador pesa un 1,6% con respecto de la comisión total.

e) Sobrefactura: Corresponde a los kilos entregados por sobrefactura, es decir, la venta generada por los transportistas bajo su gestión, sin pedido de cliente (denominada “auto venta”), la cual es pagada sólo si cumple con un nivel de servicio transportista mayor o igual a un 97%, esto es definido para que los pedidos generados por los clientes sean la prioridad de entrega.

El monto pagado corresponde al 25% de los kilos auto venta mes:

Kilos auto venta * (sector + tipo de cliente) * 25%

Este indicador corresponde al 0,6% de la comisión total.

f) Índice de complejidad de ruta: La mayor fuente de ingreso para un vendedor transportista está asociada a este indicador, ya que mide la cantidad de atenciones mes realizadas a los clientes, es decir, se realiza una cuenta de la cantidad de clientes atendidos por mes. La libre competencia que se debe generar es sólo dentro de la misma sucursal, no a nivel nacional. Este indicador varía desde un 75% como máximo hasta un mínimo de 30% del monto a total a repartir para el transportista.

Este factor se determina de la siguiente forma:

- Se determina el rango entre las atenciones de clientes (mayor número de clientes atendido por vendedor menos el menor número de clientes atendidos por vendedor).
- Se determina la desviación estándar existente entre los distintos clientes atendidos por vendedor mes.
- Luego, se divide la desviación estándar por el recorrido para obtener el intervalo de variación desde el 75% al 30%.

Tabla 6: Ejemplo participación transportistas sucursal Arica

Vendedor	Clientes atendidos	
Salazar Valenzuela Luis Ivan	918	
Pereira Castro Marco Eduardo	758	
Ubilla Martinez Francisco Eduardo	735	
Alvarado Castillo Alatiel Daniel	270	
Vega Tello Luis Hernan	73	
Olivares Henriquez Cristian Antonio	49	
	Desviación estandar	180,8
	Recorrido	869
	Variación Porcentual indice	20,8%

Este indicador representa el 15,7% de la comisión total.

g) Netomix: Este es la base de la determinación del monto a comisionar del modelo de cálculo y se realiza de la siguiente manera:

- Se toman la cantidad de kilos por mes entregados y se multiplican por la suma del sector y tipo de cliente, luego este valor es multiplicado por el porcentaje del índice de complejidad de ruta, y se le suma los kilos de sobrefactura mes por sector más tipo de cliente por un 25%, quedando la fórmula de la siguiente manera:

(kilos mes * (sector + tipo de cliente)* ICR) + (kilos sobrefactura * (sector + tipo de cliente)*25%)

Esta fórmula indica la cantidad de plata a comisionar por los kilos entregados y los kilos vendidos como sobrefactura.

h) Otros gastos: Se les paga a los transportistas todos los gastos extras que existen en la ruta asignada, como por ejemplo estacionamientos, TAG, transbordos, estadías, peonetas adicional, etc.

Este indicador corresponde al 4,8% de la comisión total.

i) Gestión del transportista: La gestión del transportista corresponde a 3 indicadores; nivel de servicio vendedor, devoluciones y precio promedio, los cuales están hechos acordes al

lineamiento financiero de la empresa. La gestión tiene un monto variable mensual máximo correspondiente a $528.000 y es la utilidad neta considerada para el transportista.

Primero que todo, se realiza una diferenciación de los tipos de transportistas existentes, ya que no todos despachan producto a los mismos clientes, por lo tanto se establecen 3 tipos de segmentos:

- Si el 70% de los kilos entregados por el transportista se concentra en supermercados, food service y/o industriales, corresponde al segmento 1.
- Si los kilos totales entregado por el transportista se concentra entre un 40% y 69% en supermercados y los demás canales de venta, es segmento 2.
- Si los kilos totales entregados por el transportista son inferior a un 39% en supermercados, food service e industriales, corresponde al segmento 3, prácticamente exclusivo de canal tradicional.

Esta segmentación permite establecer la ponderación de los indicadores de la gestión del transportista:

Tabla 7: Ponderaciones indicadores de gestión.

		Indicadores operacionales		Ventas	
Ponderaciones por segmentos		**NS Vendedor**	**Devol 4%**	**Precio promedio**	**Total**
1	**Super-Foodservice-Industriales**	**60%**	**40%**		**100%**
2	**Multiclientes**	**50%**	**20%**	**30%**	**100%**
3	**Mixto (canal tradicional)**	**40%**	**20%**	**40%**	**100%**

Cada indicador tiene el siguiente significado:

- Nivel de servicio vendedor: corresponde a la relación entre lo despachado al transportista sobre lo facturado en la ruta, es decir, que los trasnsportistas deben entregar a los clientes todos los pedidos despachados y establecidos en su hoja de ruta. La meta base de este indicador es de un 97%.
- Devolución: establece que el límite de producto a devolver a la sucursal es de un 4% con respecto a los kilos despachados mes por trasportista.
- Precio promedio: este indicador se establece sobre el 80/20 de los productos más vendidos por Agrosuper, y se mide acorde al precio promedio de la sucursal, sólo

para los sectores de pollo, cerdo y pavo, y para los clientes de canal tradicional e industriales. Este es el único indicador de la gestión que varía semanalmente y que es acorde a la realidad de cada sucursal.

Este indicador corresponde el 18,2% de la comisión total.

3. Marco teórico

Para el desarrollo de esta tesis, se analizan los datos con el propósito de descubrir, extraer y almacenar información relevante de las base de datos de las comisiones de los transportistas de Agrosuper, a través de programas de búsqueda e identificación de patrones y relaciones globales, tendencias, desviaciones y otros indicadores, que tienen una explicación y que pueden descubrirse mediante diversas técnicas de las herramientas utilizadas.

3.1. Tratamiento y procesamiento de datos

El aumento del volumen y variedad de información que se encuentra informatizada en bases de datos digitales ha crecido espectacularmente en la última década. Gran parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido en el tiempo, la cual además de ser la "memoria de la organización", es útil para predecir la información futura.

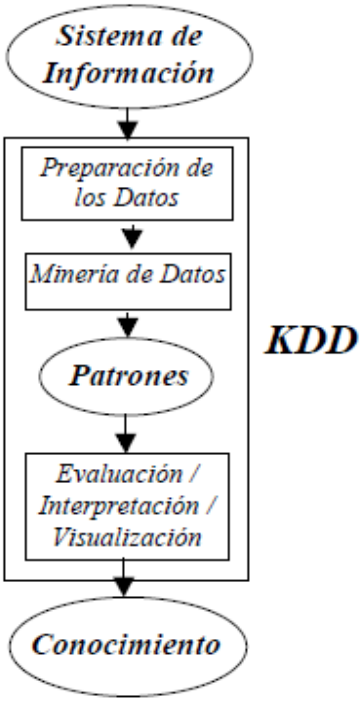
En base a esto, surgen nuevas necesidades de análisis de grandes volúmenes de información, y una de los métodos utilizados para esto es el "Descubrimiento de Conocimiento a partir de Bases de Datos" (*KDD,* del inglés *Knowledge Discovery from Databases*), el cual se define de la siguiente manera:

La metodología de KDD cambia la manera de extraer el conocimiento, se hace más eficiente, permite tener entornos de descubrimiento ('navegación') y genera consultas inductivas. Esta metodología nace como una interfaz y se nutre de diferentes disciplinas: estadística, sistemas de información / bases de datos, aprendizaje automático-IA, visualización de datos, computación paralela-distribuida e interfaces de lenguaje natural a bases de datos.

Las fases que componen esta metodología son las siguientes:

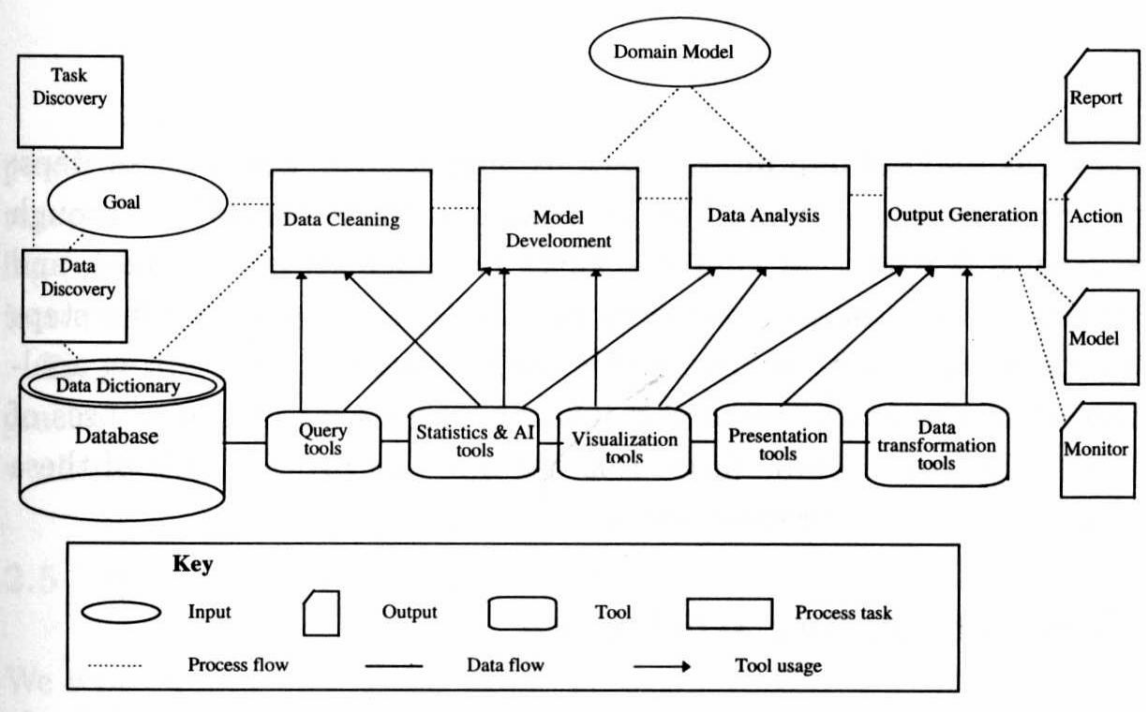
a) Determinación de las fuentes de información que pueden ser útiles y establecer dónde conseguirlas.
b) Diseño del esquema de un almacén de datos que consiga unificar de manera operativa toda la información recogida.
c) Constitución del almacén de datos que permita la navegación y visualización previa de la información, para discernir qué aspectos puede interesar que sean estudiados.
d) Selección, limpieza y transformación de los datos que se van a analizar.
e) Seleccionar y aplicar el método de minería de datos apropiado.
f) Interpretación, transformación y representación de los patrones extraídos.
g) Difusión y uso del nuevo conocimiento.

Figura 4: Flujo de metodología KDD



Además, las distintas técnicas de distintas disciplinas se utilizan en diferentes fases:

Figura 5: Uso de diferentes técnicas en la metodología KDD.



El levantamiento de datos, en las primeras fases del KDD, determina que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original. Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra en bases de datos y otras fuentes muy diversas, tanto internas como externas.

La limpieza de datos (data cleaning) y criba (selección) tiene como primer paso la elaboración de un resumen con las características de los mismos, en donde se debe eliminar el mayor número posible de datos erróneos o inconsistentes (limpieza) e irrelevantes (criba). Para esto, se utilizan casi exclusivamente métodos estadísticos tales como histogramas (detección de datos anómalos), selección de datos y redefinición de atributos, dentro de los cuales existen los siguientes:

- Atributos Nominales: para la detección de valores redundantes y despreciables.
- Atributos Numéricos: para la detección de valores anómalos y distribuciones en los datos.

3.2. Minería de datos

Las empresas acumulan grandes cantidades de datos que son almacenados, algunos de ellos son utilizados, otros se acumulan hasta perderse por falta de actualización o por cambios de políticas de manejo. Ahora con el desarrollo de sistemas de cómputo, las empresas tienen la capacidad de almacenar y accesar, en archivos o bases de datos, grandes cantidades de datos históricos sobre las operaciones de su negocio; información que en su momento fue usada para satisfacer las necesidades propias de la empresa y como soporte de las decisiones. Todos esos archivos contienen normalmente gran cantidad de datos que serían de utilidad si fuera posible aprovecharlos mediante procesos que arrojaran información útil.

Las áreas de sistemas han trabajado en la creación de extractores de información de las bases de datos operacionales y en el almacenamiento de estos datos en archivos, tratando de responder a las peticiones de los usuarios que necesiten obtener información que les ayude a tomar mejores decisiones. Las necesidades de información han hecho que se diseñen sistemas de información ejecutiva y de apoyo a la toma de decisiones, sin embargo, las demandas de las empresas, con relación a la información, van más allá de simples consultas, tabulaciones cruzadas o reportes consolidados; lo que ha hecho que se creen nuevas formas de análisis de la información, con ventajas respecto de las que se conocían porque incorporan hechos sistemáticos que relacionan más de dos variables.

La minería de datos es el proceso que tiene como propósito descubrir, extraer y almacenar información relevante de amplias bases de datos, a través de programas de búsqueda e identificación de patrones y relaciones globales, tendencias, desviaciones y otros indicadores aparentemente caóticos que tienen una explicación que se pueden descubrir mediante diversas técnicas de esta herramienta.

El objetivo fundamental es aprovechar el valor de la información localizada y usar los patrones preestablecidos para poder tener un mejor conocimiento del negocio y así tomar decisiones más confiables.

3.2.1 Ventajas

- Apoyar en el procesamiento de datos para descubrir relaciones que eran desconocidas.
- Permitir elegir cursos de acción y definir estrategias competitivas.
- Inferir relaciones en grandes volúmenes de datos, mediante modelos avanzados y reglas de inducción, ya que puede examinar gran cantidad de datos y encontrar patrones difíciles de identificar a simple vista.
- Puede trabajar siguiendo los mismos criterios con grandes cantidades de información histórica.
- El proceso de búsqueda puede ser realizado por herramientas que automáticamente buscan patrones porque así están programadas y despliegan los tópicos más importantes.

3.2.2 Estructura

- Algortimos o programa de búsqueda mineros

Se hace uso de programas de búsqueda para detectar desviaciones, tendencias y patrones ocultos en los datos históricos.

Los mineros son programas pensados y creados por el usuario, en los que se emplean técnicas diferentes para la explotación de datos, tales como cluster, asociaciones, clasificación, visualización, redes neuronales, algoritmos genéticos, detección de desviaciones, entre otros. Todos ellos requieren bases de datos de tamaño considerable para que puedan ser eficientes.

La función de los programas mineros es correlacionar los criterios de selección y búsqueda con los datos históricos y cuando encuentran algo interesante es presentado como un hallazgo.

Los programas mineros trabajan con procesos automáticos principalmente, sobre bases de dato, patrones, tendencias o desviaciones; una de las ventajas de los mineros es que no requieren hardware especial o dedicado.

- Algoritmos de aprendizaje de árboles de decisión escalables

Son un modelo predictivo utilizado en el ámbito de la inteligencia artificial y el análisis predictivo, dada una base de datos se generan estos diagramas de decisiones lógicas que sirven para representar y categorizar una serie de condiciones que suceden de forma continua, para la resolución de un problema.

Son diseñados bajo los requerimientos de que no sea necesario que los datos quepan en memoria y los chequeos de consistencia se realicen eficientemente, utilizando índices, con el objetivo de agilizar los escaneos de los datos:

a) Datos históricos (en dónde buscan): Datos estables y coherentes que se van acumulando a lo largo de la vida operativa de una empresa.
b) Criterios de búsqueda (qué busca): Normas, tendencias y patrones desde los cuales los programas mineros realizarán el proceso de selección y búsqueda en los datos históricos. La prioridad de búsqueda, los criterios de interés y las explicaciones de situaciones extrañas son definidos por el usuario. Una vez establecidos los criterios de selección y búsqueda se analizan los datos históricos reportando los hallazgos inmediatamente en un archivo para su posterior revisión y decisión final.
c) Almacenamiento de hallazgos: Los hallazgos son los datos resultantes de correlacionar los criterios de selección y búsqueda con los datos históricos. El usuario desempeña un papel fundamental, ya que es él quien puede decidir si este patrón, tendencia o criterio, tiene importancia, pertinencia y utilidad.

3.2.3 Ciclo de minería de datos

El proceso de la minería de datos es un ciclo, debido a que los resultados obtenidos pueden alimentar nuevamente dicho proceso; intervienen, principalmente, cuatro pasos que se describen a continuación 3:

a) Los usuarios de la información deberán identificar los problemas del negocio y las áreas en donde los datos pueden dar valor agregado a la empresa, esto quiere decir que, a través de un programa surge la necesidad de analizar a detalle los datos de la empresa para poder encontrar posibles soluciones al mismo, o bien, información que haga que las decisiones tomadas sean lo más certeras posibles. Igualmente, es importante identificar las áreas en donde la información es cambiante, pero primordial para la competitividad de

la empresa. Para esto pueden manejarse diferentes criterios, no se puede decir específicamente cuáles son los correctos debido a que esto depende de las características de la empresa, pero el objetivo a perseguir es determinar los criterios, ideas, normas y cuestionamientos que fungirán como entrada para el proceso de minería de datos.

Para analizar la información histórica se debe seleccionar el algoritmo o algoritmos adecuados de minería. Posteriormente, estos algoritmos son traducidos a programas mineros que realizarán las búsquedas con los criterios previamente definidos.

b) Incorporar la información obtenida a través del proceso de minería de datos al proceso de toma de decisiones; así como presentar los hallazgos encontrados a los responsables de las operaciones de forma que la información obtenida pueda integrarse en los procesos de la empresa y pueda aplicarse en la solución de los problemas.
c) Medir los resultados: cuantificar el valor de los hallazgos encontrados en relación a la solución de los problemas identificados y a los criterios definidos en el primer punto.

Actualmente el valor de la información se ha acrecentado hasta convertirse en un activo estratégico para la competitividad de una empresa. Su unidad y consistencia son importantes, pues de estas características depende una buena parte de la confiabilidad de la información seleccionada para tomar decisiones.

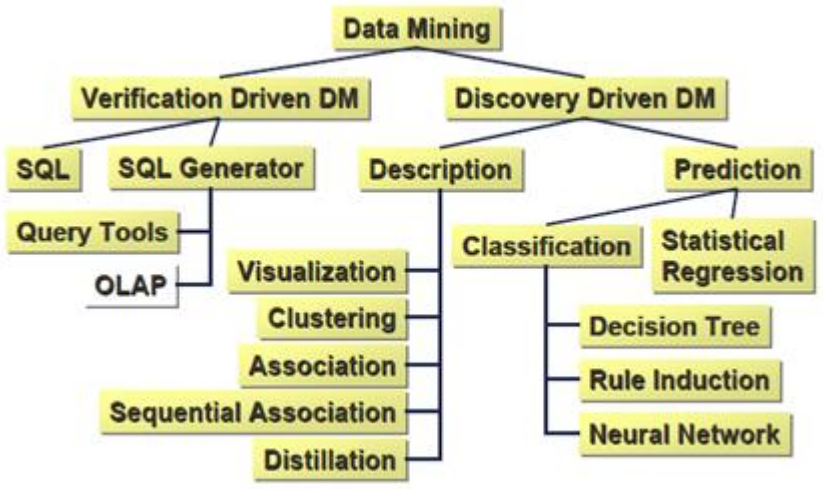
La minería de datos ayuda a obtener una visión más completa y detallada del negocio, ya que permite buscar datos de operaciones cotidianas que se salen de los rangos que están considerados como normales de lo que, en parte, depende de la confiabilidad de la información para la toma de decisiones.

En la medida en que una empresa capte datos de sus operaciones cotidianas tendrá la oportunidad de correlacionarlos y hacer descubrimientos que le ayuden a identificar posibles clientes, puntos de venta, fraudes, entre otros.

La minería de datos tiene futuro dentro de las empresas, debido a que existen grandes bases de datos que contienen valores desaprovechados; los mercados están más saturados y se requieren de análisis intensos para captar la atención de los clientes.

En todo el proceso de la minería de datos, el ser humano es el factor más importante, ya que sólo él tiene la capacidad de analizar y decidir si los patrones, normas o funciones encontrados tienen importancia, pertinencia y utilidad para su empresa.

Figura 6: Taxonomía de las técnicas de minería de datos



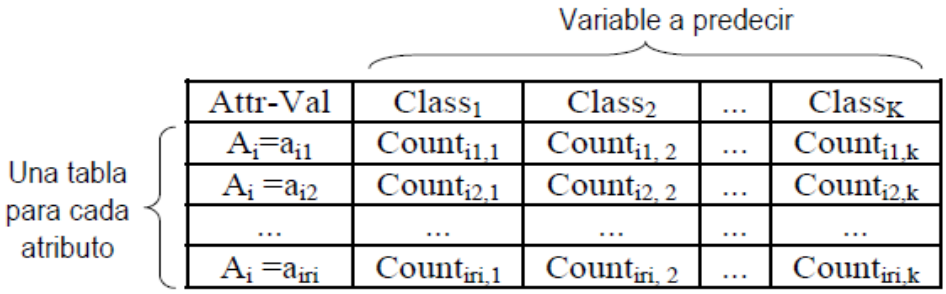
3.2.3.1 Métodos específicos minería de datos

a) Algoritmos de aprendizaje de árboles de decisiones.

Se diseñan bajo los siguientes requerimientos:

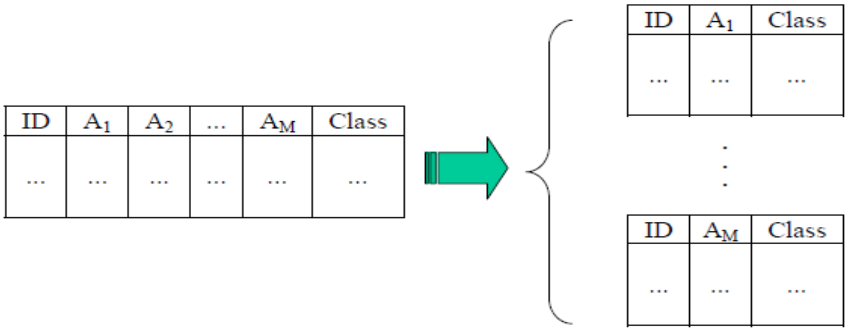
- No debe ser necesario que los datos sean almacenados en memoria.
- Los chequeos de consistencia se hacen eficientemente, utilizando índices con el objetivo de agilizar el escaneo de los datos.
- Las condiciones sobre índices son preferibles sobre aquellas que no permiten indización.
- Se utilizan las cuentas de correlación:

Figura 7: Tabla para cuentas de correlación.



- Se puede realizar una partición vertical de las tablas mencionadas anteriormente:

Figura 8: Ejemplo de partición vertical de tabla.



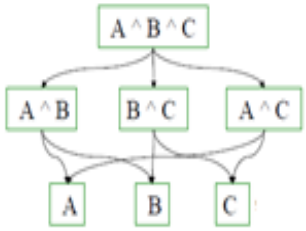
b) Dependencias Funcionales

$$A \wedge B \wedge C \rightarrow D$$

Significa que para los mismos valores de A, B y C existe un mismo valor de D, es decir, es función de A, B y C.

Si se representa la parte izquierda como un conjunto de condiciones, se puede establecer una relación de orden entre las dependencias funcionales, lo cual genera un semi-retículo.

Figura 9: Representación del semi-retículo.



c) Correlaciones y estudios factoriales: Permite establecer la relevancia e irrelevancia de los factores, y si ésta es positiva o negativa con respecto a otro factor. A continuación se ejemplificará lo anteriormente mencionado:

Estudio de visitas: 11 pacientes, 7 factores:

• Health: salud del paciente (referida a la capacidad de ir a la consulta). (1-10)

• Need: convicción del paciente que la visita es importante. (1-10)

• Transportation: disponibilidad de transporte del paciente al centro. (1-10)

• Child Care: disponibilidad de dejar los niños a cuidado. (1-10)

• Sick Time: si el paciente está trabajando, puede darse de baja. (1-10)

• Satisfaction: satisfacción del cliente con su médico. (1-10)

• Ease: facilidad del centro para concertar cita y eficiencia de la misma. (1-10)

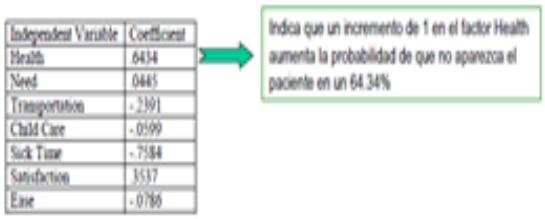
• No-Show: indica si el paciente no se ha pasado por el médico durante el último año (0-se ha pasado, 1 no se ha pasado)

Matriz de correlaciones:

Tabla 8: Tabla ejemplo descrito

	Health	Need	Transp'tion	Child Care	Sick Time	Satisfaction	Ease	No-Show
Health	1							
Need	-0.7378	1						
Transportation	0.3116	-01041	1					
Child Care	0.3116	-01041	1	1				
Sick Time	0.2771	0.0602	0.6228	0.6228	1			
Satisfaction	0.22008	-0.1337	0.6538	0.6538	0.6257	1		
Ease	0.3887	-0.0334	0.6504	0.6504	0.6588	0.8964	1	
No-Show	0.3955	-0.5416	-0.5031	-0.5031	-0.7249	-0.3988	-0.3278	1

Coeficientes de regresión:



Independent Variable	Coefficient
Health	.6434
Need	.0445
Transportation	- .2391
Child Care	- .0599
Sick Time	- .7584
Satisfaction	.3537
Ease	- .0786

d) Multi-Clasificadores (clustering): Para obtener resultados en la clasificación es necesario tener un conjunto de modelos, denominados *Ensembles.* La metología para generar esto es la siguiente:
    - Manipulación de los datos de entrenamiento
    - Manipulación de los atributos
    - Manipulación de las clases
    - Métodos aleatorios.

4. Weka

Weka es un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos. Es un software que ha sido desarrollado bajo licencia GPL lo cual ha impulsado que sea una de las suites más utilizadas en el área en los últimos años. Incluye las siguientes características:

a) Diversas fuentes de datos (ASCII, JDBC).
b) Interfaz visual basado en procesos/flujos de datos (rutas).
c) Distintas herramientas de minería de datos: reglas de asociación, agrupación/segmentación/conglomerado, clasificación y regresión.
d) Manipulación de datos (pick & mix, muestreo, combinación y separación).
e) Combinación de modelos.
f) Visualización anterior (datos en múltiples gráficas) y posterior (árboles, curvas ROC, curvas de coste, etc.).
g) Entorno de experimentos, con la posibilidad de realizar pruebas estadísticas.

Figura 10: Ventana inicial de weka

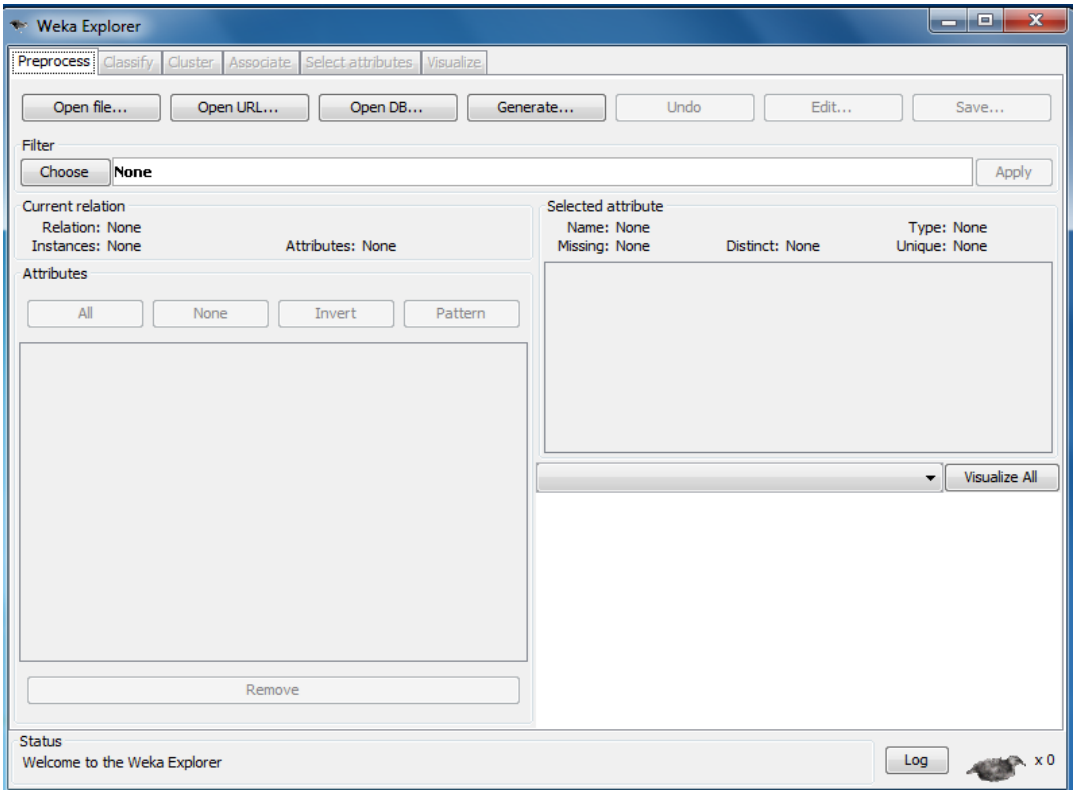
Como se ve en la parte derecha de la Figura 10, Weka define 4 entornos de trabajo:

- Explorer: Entorno visual que ofrece una interfaz gráfica para el uso de los paquetes.
- Experimenter: Entorno centrado en la automatización de tareas de manera que se facilite la realización de experimentos a gran escala.
- KnowledgeFlow: Permite generar proyectos de minería de datos mediante la generación de flujos de información.
- Simple CLI: Entorno consola para invocar directamente con Java los paquetes de Weka.

4.1 Explorer

El modo Explorador es el más usado y descriptivo. Permite realizar operaciones sobre un sólo archivo de datos y permite el acceso a la mayoría de las funcionalidades integradas en Weka de una manera sencilla.

Figura 11: Ventana del explorador.



Se observa que existen 6 sub-entornos de ejecución:

- Preprocess: Incluye las herramientas y filtros para cargar y manipular los datos.
- Classification: Acceso a las técnicas de clasificación y regresión.
- Cluster: Integra varios métodos de agrupamiento.
- Associate: Incluye una pocas técnicas de reglas de asociación.
- Select Attributes: Permite aplicar diversas técnicas para la reducción del número de atributos.
- Visualize: En este apartado se puede estudiar el comportamiento de los datos mediante técnicas de visualización.

4.1.1 Preprocess

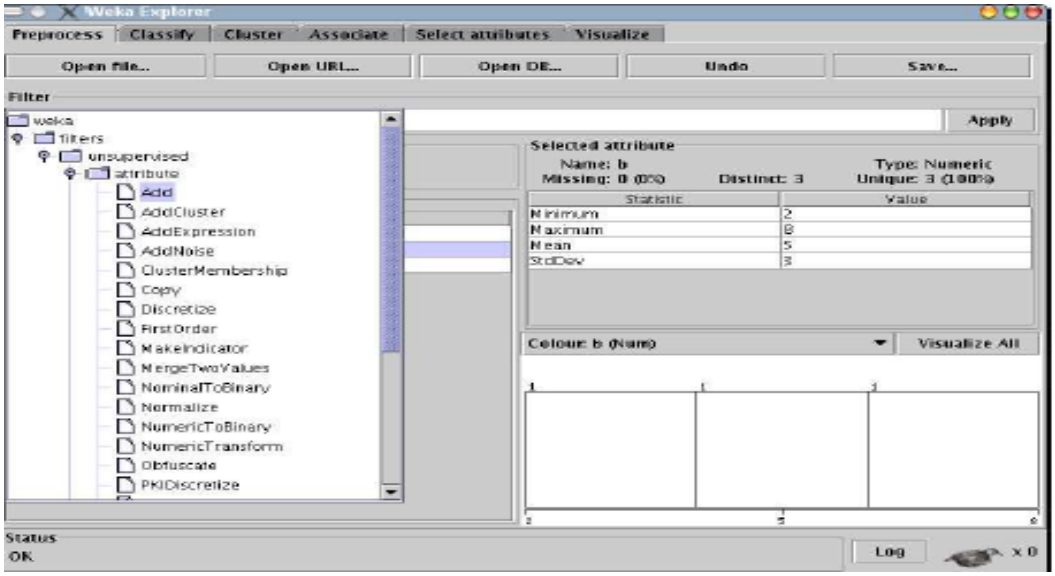
El primer paso para comenzar a trabajar con el explorador es definir el origen de los datos. Weka soporta diferentes fuentes que coinciden con los botones que están debajo de las pestañas superiores mostrados en la figura 4. Las diferentes posibilidades son las siguientes:

- Open File: al pulsar sobre este botón aparecerá una ventana de selección de fichero.
- Open Url: con este botón se abrirá una ventana que permite introducir una dirección en la que definir dónde se encuentra el fichero.
- Open DB: este botón da la posibilidad de obtener los datos de una base de datos.

4.1.1.1 Aplicación de filtros

Weka permite aplicar una gran diversidad de filtros sobre los datos, permitiendo realizar transformaciones sobre ellos de todo tipo.

Figura 12: Aplicación de un filtro en el modo Explorador.



A continuación se hace una breve descripción de los filtros aplicados en la categoría: a) Atributte

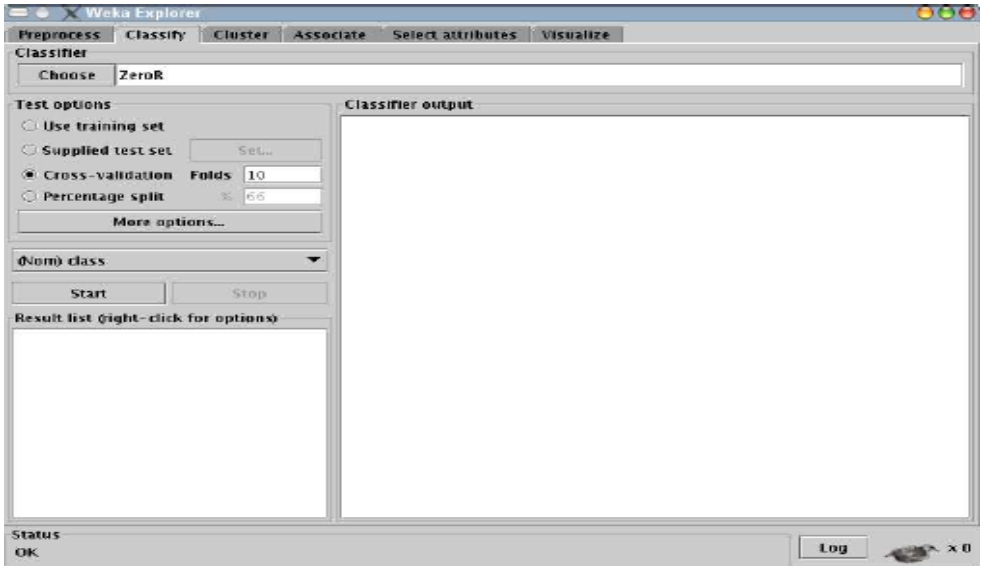
- *Add*: Añade un atributo más.
- *Add Expression*: permite agregar un atributo que sea el valor de una función.
- *Add Noise*: añade ruido a un determinado atributo que debe ser nominal
- *ClusterMember*: filtro que dado un conjunto de atributos y el atributo que define la clase de los mismos, devuelve la probabilidad de cada uno de los atributos de estar clasificados en una clase u otra.
- *Copy*: realiza una copia de un conjunto de atributos en los datos.
- *Discretize*: discretiza un conjunto de valores numéricos en rango de datos.
- *FistOrder*: Este filtro realiza una transformación de los datos obteniendo la diferencia de pares consecutivos de datos, suponiendo un dato inicial adicional de valor 0 para conseguir que la cardinalidad del grupo de datos resultante sea la misma que la de los datos origen.
- *MakeIndicator*: crea un nuevo conjunto de datos reemplazando un atributo nominal por uno booleano (asignará “1” si en una instancia se encuentra el atributo nominal seleccionado y “0” en caso contrario).
- *MergeTwoValues*: fusiona dos atributos nominales en uno solo.
- *NominalToBinary*: transforma los valores nominales de un atributo en un vector cuyas coordenadas son binarias.

- *Normalize*: normaliza todos los datos de manera que el rango de los datos pase a ser [0,1].
- *NumericToBinary*: convierte datos en formato numérico a binario.
- *NumericTransform*: Filtro similar a AddExpression pero mucho más potente. Permite aplicar un método java sobre un conjunto de atributos dándole el nombre de una clase y un método.
- *Obfuscate*: ofusca todas las cadenas de texto de los datos.
- *PKIDiscretize*: discretiza atributos numéricos (al igual que Discretize), pero el número de intervalos es igual a la raíz cuadrada del número de valores definidos.
- *RandomProjection*: reduce la dimensionalidad de los datos.
- *Gaussian*: utiliza una distribución gaussiana.
- *Remove*: borra un conjunto de atributos del fichero de datos.
- *RemoveType*: elimina el conjunto de atributos de un tipo determinado.
- *RemoveUseless*: elimina atributos que oscilan menos que un nivel de variación.
- *ReplaceMissingValues*: reemplaza todos los valores indefinidos por la moda en el caso de que sea un atributo nominal o la media aritmética si es un atributo numérico.
- *Standarize*: estandariza los datos numéricos de la muestra para que tengan de media 0 y la unidad de varianza.
- *StringToNominal*: convierte un atributo de tipo cadena en un tipo nominal.
- *StringToWordVector***:** convierte los atributos de tipo String en un conjunto de atributos representando la ocurrencia de las palabras del texto.
- SwapValues: intercambia los valores de dos atributos nominales.
- *TimeSeriesDelta*: filtro que asume que las instancias forman parte de una serie temporal y reemplaza los valores de los atributos de forma que cada valor de una instancia es reemplazado con la diferencia entre el valor actual y el valor pronosticado para dicha instancia.
- *Instance*: los filtros son aplicados a instancias concretas enteras.

4.1.2 Classify (clasificación)

Pulsando en la segunda pestaña (zona superior) del explorador se ingresa en el modo clasificación (figura 6). En este modo se puede clasificar por varios métodos los datos ya cargados.

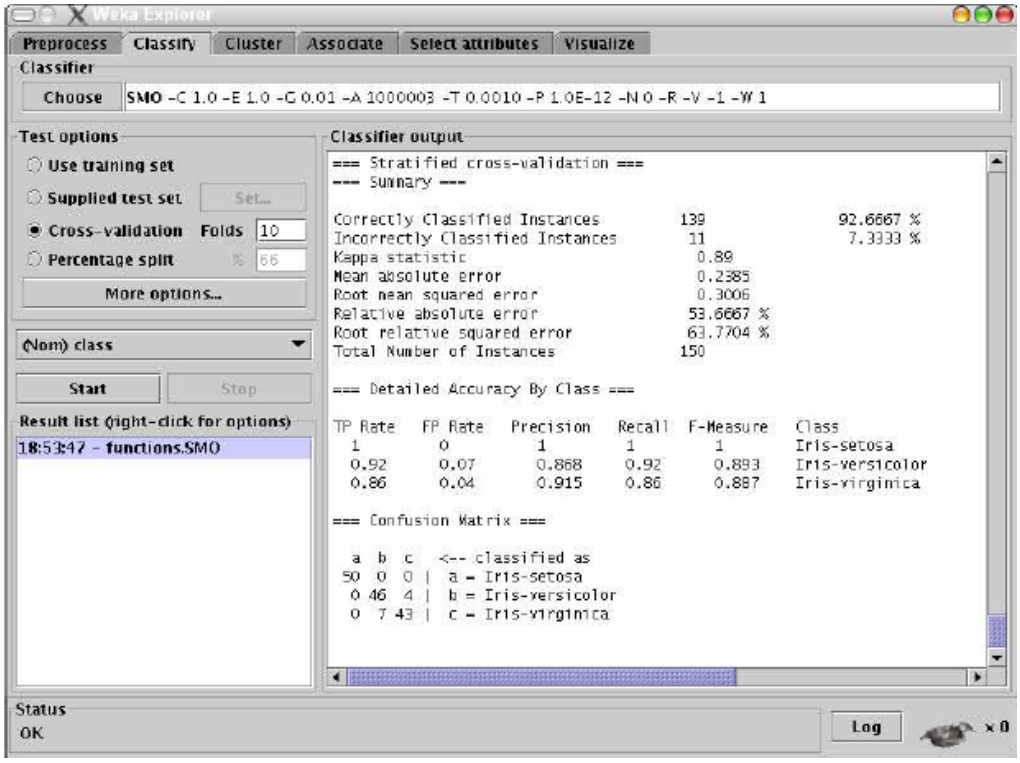
Figura 13: Modo clasificación dentro del explorador.



Para realizar una clasificación se debe elegir un clasificador y configurarlo al requerimiento del usuario. Una vez elegido el clasificador y sus características se debe configurar el modo de entrenamiento (Test Options). Weka proporciona 4 modos de prueba:

- *Use training set*: con esta opción Weka entrenará el método con todos los datos disponibles y luego lo aplicará otra vez sobre los mismos.
- *Supplied test set*: marcando esta opción se puede seleccionar un fichero de datos con el que se probará el clasificador obtenido con el método de clasificación usado y los datos iniciales.
- *Cross-validation*: Weka realiza una validación cruzada estratificada del número de particiones dado (Folds).
- *Percentage Split*: se define un porcentaje con el que se construirá el clasificador y con la parte restante se probará.

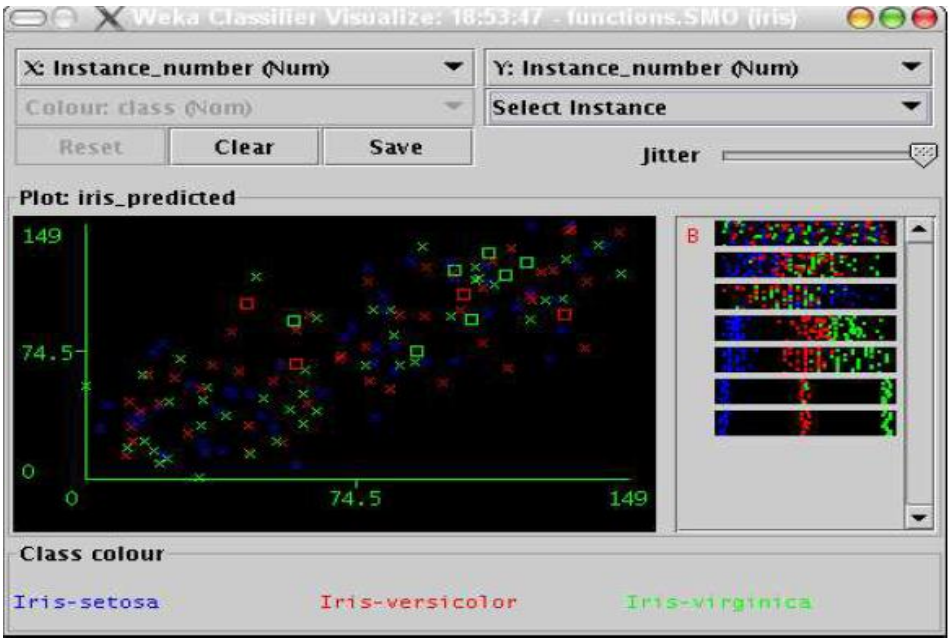
Figura 14: Aplicación de un método de clasificación.



Además, en Clasificación se puede ver el resultado del experimento de las siguientes maneras:

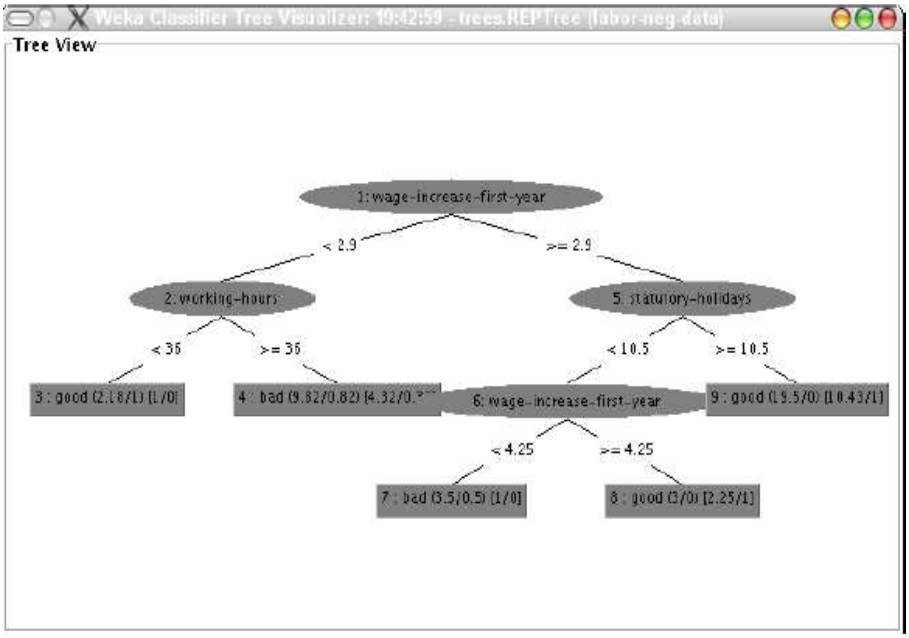
- *View in main window*: se muestra el resultado del experimento en la salida estándar del clasificador.
- *View in separate window*: se muestra el resultado del experimento en una nueva ventana.
- *Save result buffer*: se guarda el resultado del experimento en un fichero.
- *Load model*: se carga un modelo de clasificador ya construido.
- *Save model*: se guarda el modelo de clasificador actual.
- *Re-evaluate model on current test set*: enfrenta un modelo con el conjunto de muestra actual.
- *Visualize clasiffier errors*: se abre una nueva ventana en la que nos mostrará una gráfica con los errores de clasificación.

Figura 15: Gráfica de los errores de Clasificación.



- *Visualize tree*: esta opción muestra un árbol de decisión, como el de la figura 9, generado por el clasificador.

Figura 16: Visualización de árboles de decisión.

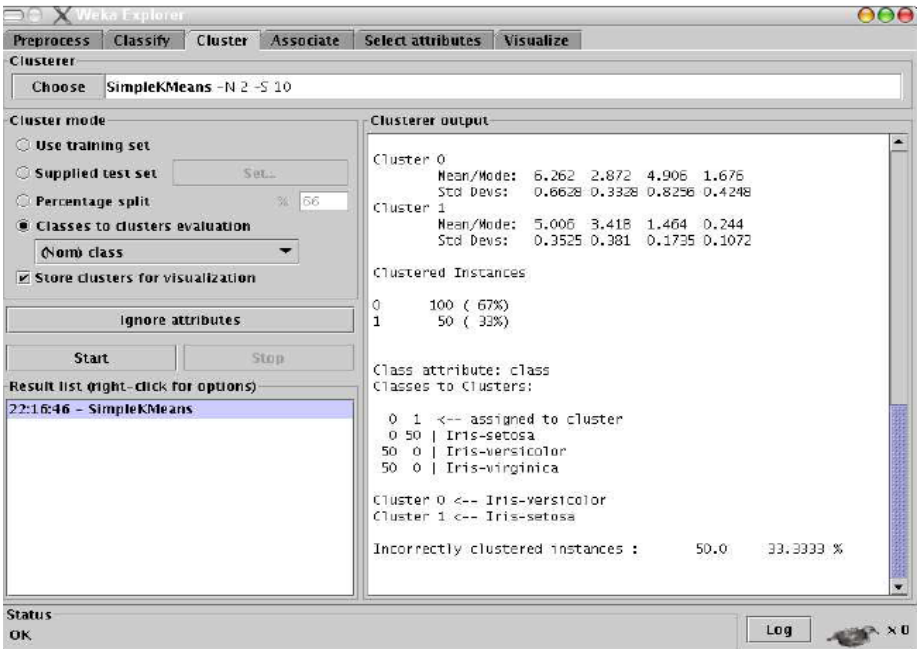


- *Visualize margin curve*: muestra en una curva la diferencia entre la probabilidad de la clase estimada y la máxima probabilidad de otras clases.
- *Visualize thresold curve*: muestra la variación de las proporciones de cada clase.
- *Visualize cost curve*: muestra una gráfica que indica la probabilidad de coste al variar la sensibilidad entre clases.

4.1.3 Cluster

Una opción propia de este apartado es la posibilidad de ver de una forma gráfica la asignación de las muestras en clusters.

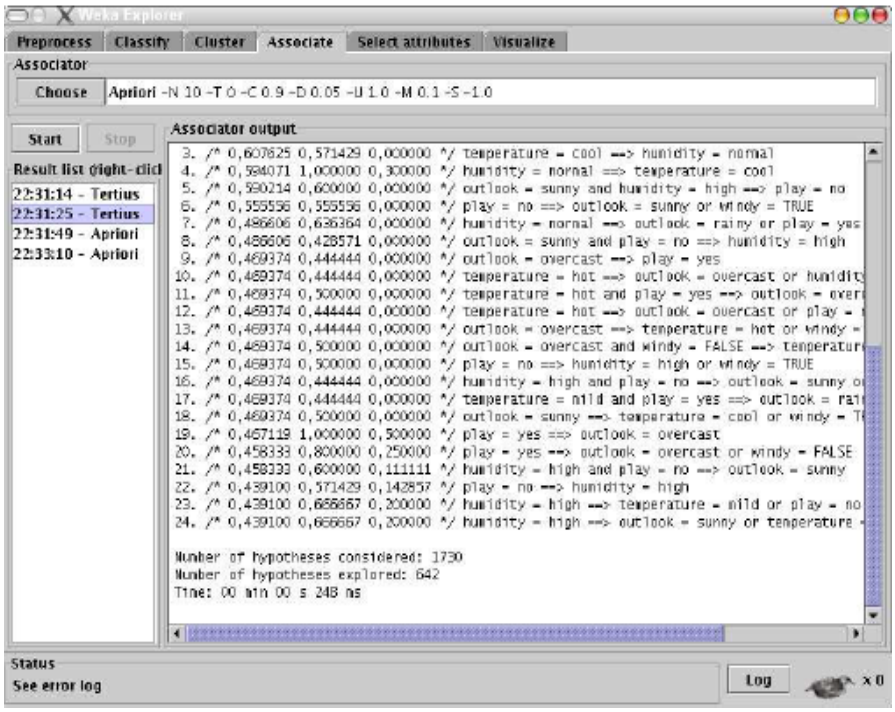
Figura 17: Modo cluster dentro del modo explorador.



4.1.4 Associate (búsqueda de asociaciones)

Permite aplicar métodos orientados a buscar asociaciones entre datos. Es importante señalar que estos métodos sólo funcionan con datos nominales.

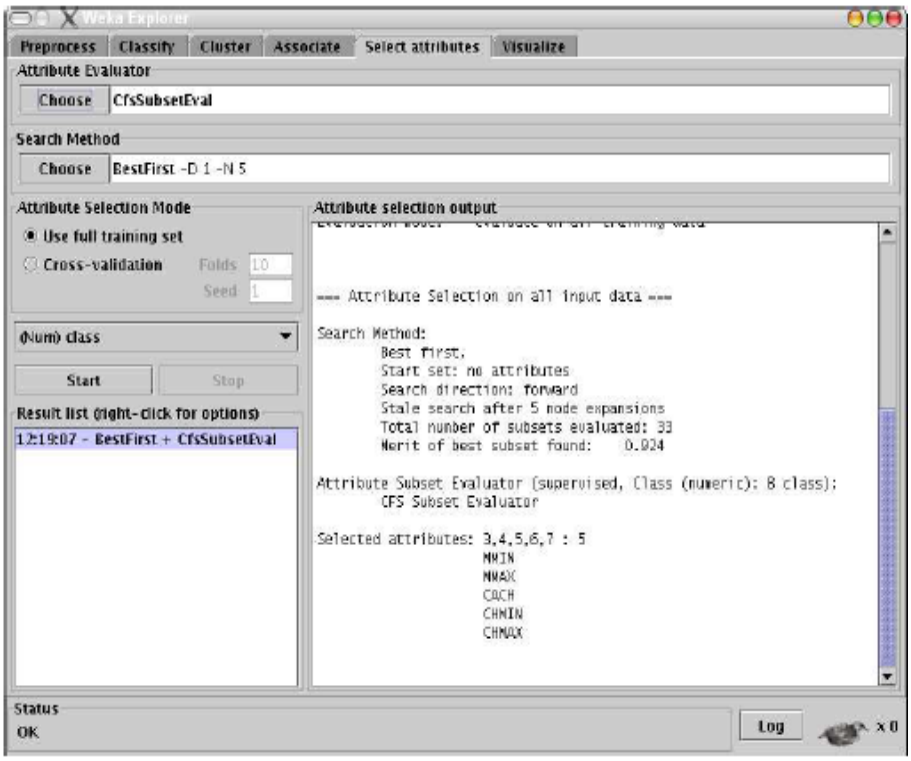
Figura 18: Modo asociación dentro del modo explorador.



4.1.5 Select attributes (selección de atributos)

El objetivo de estos métodos es identificar, mediante un conjunto de datos que poseen ciertos atributos, aquellos atributos que tienen más peso a la hora de determinar si los datos son de una clase u otra.

Figura 19: Modo de selección de atributos dentro del modo explorador.

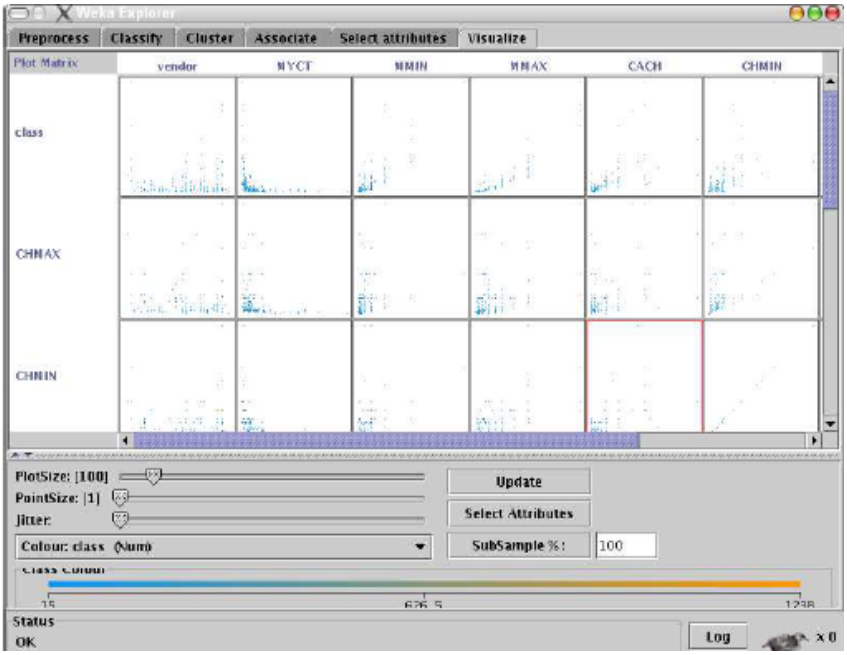


Este método es el encargado de evaluar cada uno de los casos a los que se le enfrente y dotar a cada atributo de un peso específico.

4.1.6 Visualize (visualización)

Es un modo que muestra gráficamente la distribución de todos los atributos mostrando gráficas en dos dimensiones, en las que va representando en los ejes todos los posibles pares de combinaciones de los atributos. Este modo permite ver correlaciones y asociaciones entre los atributos de una forma gráfica.

Figura 20: Modo de Visualización dentro del modo explorador.



4.2 Experimenter (experimentador)

El modo experimentador (*Experimenter*) es un modo muy útil para aplicar uno o varios métodos de clasificación sobre un gran conjunto de datos y, luego poder realizar contrastes estadísticos entre ellos y obtener otros índices estadísticos.

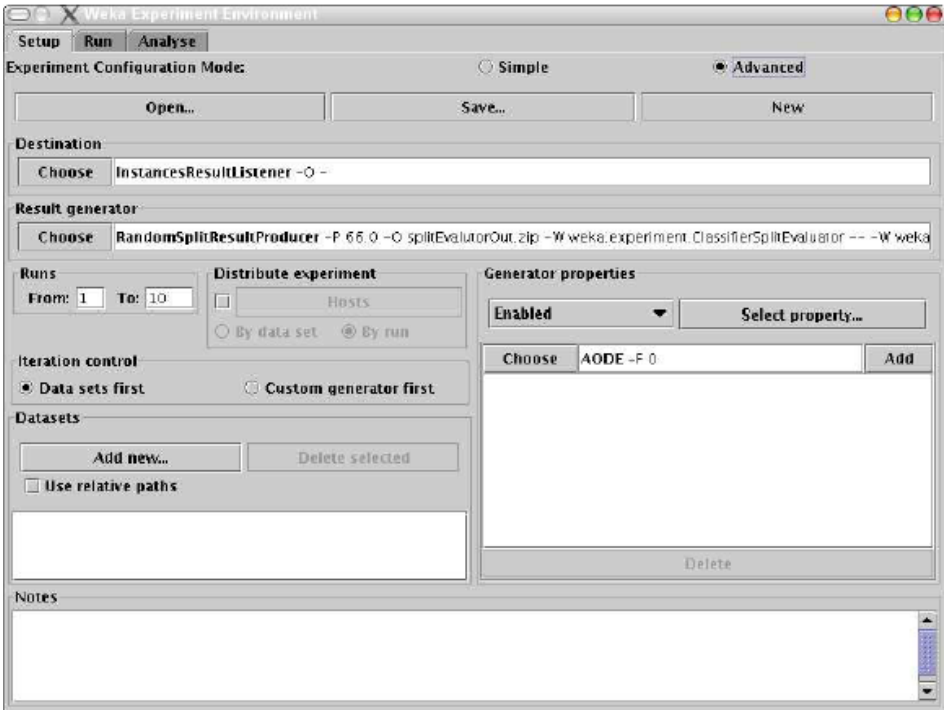
Figura 21: Modo experimentador, modo simple.

Al igual que en el modo Explorer, existen tres tipos de validaciones: validación-cruzada estratificada, entrenamiento con un porcentaje de la población tomando ese porcentaje de forma aleatoria y entrenamiento con un porcentaje de la población tomando el porcentaje de forma ordenada.

El comando Iteration control define el número de repeticiones del experimento, especificando si se deben realizar primero los archivos de datos o los algoritmos.

A su vez, el modo Experimentador cuenta con un modo avanzado (figura 22):

Figura 22: Modo experimentador, modo advanced.



La principal diferencia es que el funcionamiento de este modo está orientado a realizar tareas específicas más concretas que un experimento normal, y una cierta funcionalidad existente en el modo simple se ha traslado al modo avanzado, mostrándola más concreta y explícita al usuario.

El cambio fundamental entre una interfaz y otra se basa en el *Result Generator*. En este nuevo modo es necesario seleccionar qué método generador de resultados se utilizará y seguidamente configurarlo. Los 5 métodos que permite seleccionar Weka son:

a) *CrossValidationResultProducer*: genera resultados fruto de una validación cruzada. Tiene por opciones:
   - *Numfolds*: el número de particiones para la validación cruzada.
   - *OutputFile*: fichero donde se guarda cada una de las particiones que se realizarán en la validación cruzada. Lo guarda codificado en formato Zip.
   - *RawOutput*: guarda los resultados "tal cual", es decir sin ningún formato fijo.
   - *SplitEvaluator*: el método que se selecciona para evaluar cada una de las particiones de la validación cruzada. En él se especifica el clasificador a usar.

b) *AveragingResultProducer*: toma los resultados de un método generador de resultados y calcula los promedios de los resultados. Tiene las siguiente opciones:
   - *CalculateStdDevs*: calcula las desviaciones estándar.
   - *ExpectedResultsPerAverage*: elige el número de resultados que se esperan, para dividir la suma de todos los resultados entre este número.
   - *KeyFieldname*: se selecciona el campo que será distintivo y único de cada repetición. Por defecto es "Fold" ya que cada repetición será de una partición distinta.
   - *ResultProducer*: el método generador de resultados del que tomará los datos.

c) *LearningRateResultProducer*: método generador de resultados para ir repitiendo el experimento variando el tamaño del conjunto de datos. Habitualmente se usa con un AveragingResultProducer y CrossValidationResultProducer para generar curvas de aprendizaje. Sus opciones son:
   - *Lowersize*: selecciona el número de instancias para empezar.
   - *ResultProducer*: selecciona el método generador de resultados.
   - *Setsize*: número de instancias a añadir por iteración.
   - *UpperSize*: selecciona el número máximo de instancias en el conjunto de datos.

d) *RandomSplitResultProducer*: genera un conjunto de entrenamiento y de prueba aleatorio para un método de clasificación dado.
   - *Outputfile* define el archivo donde se guardarán los resultados.
   - *RandomizeData*: si se activa, toma los datos de forma aleatoria.
   - *RawOutput*: proporciona resultados sin formato útiles para depurar el programa.
   - *SplitEvaluator*: selecciona el método de clasificación a usar.

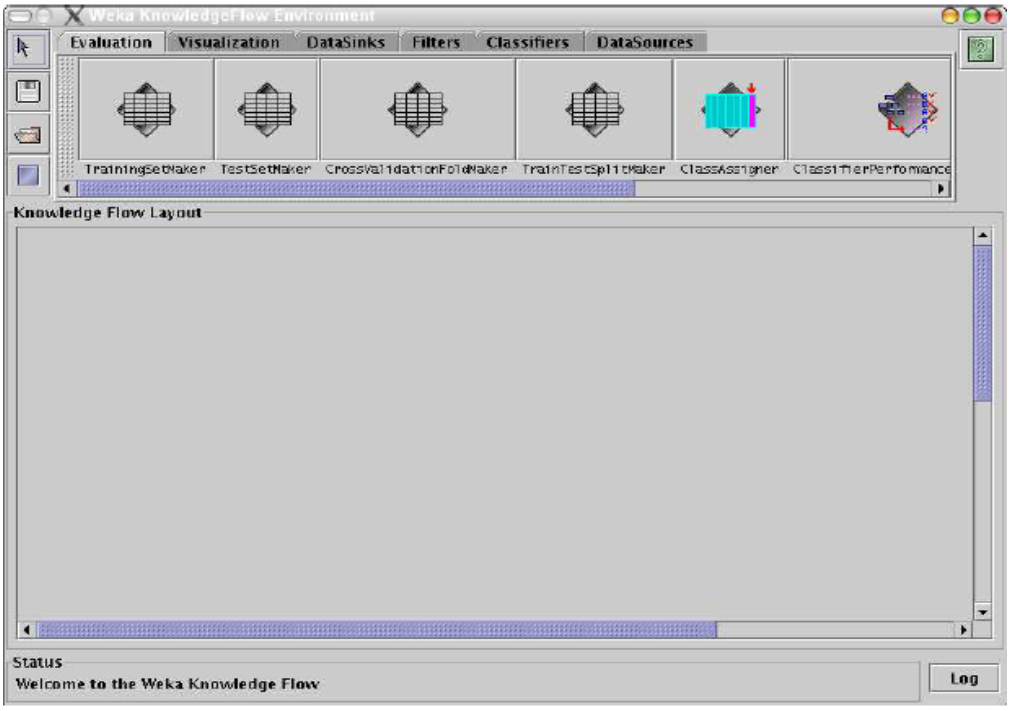
- *TrainPercent*: establece el porcentaje de datos para entrenar la muestra.

e) *DatabaseResultProducer*: de una base de datos toma los resultados que coinciden con los obtenidos con un método generador de resultados (cualquiera de los anteriores). Si existen datos que no aparezcan en la base de datos se calculan.

4.3 Knowledge Flow

Esta interface de Weka muestra de una forma más explícita el funcionamiento interno del programa. Su funcionamiento es gráfico y se basa en situar en el panel de trabajo (zona gris de la figura 16), elementos base (situados en la parte superior de la ventana) de manera que se cree un “circuito” que defina el experimento.

Figura 23: Modo knowledge flow.

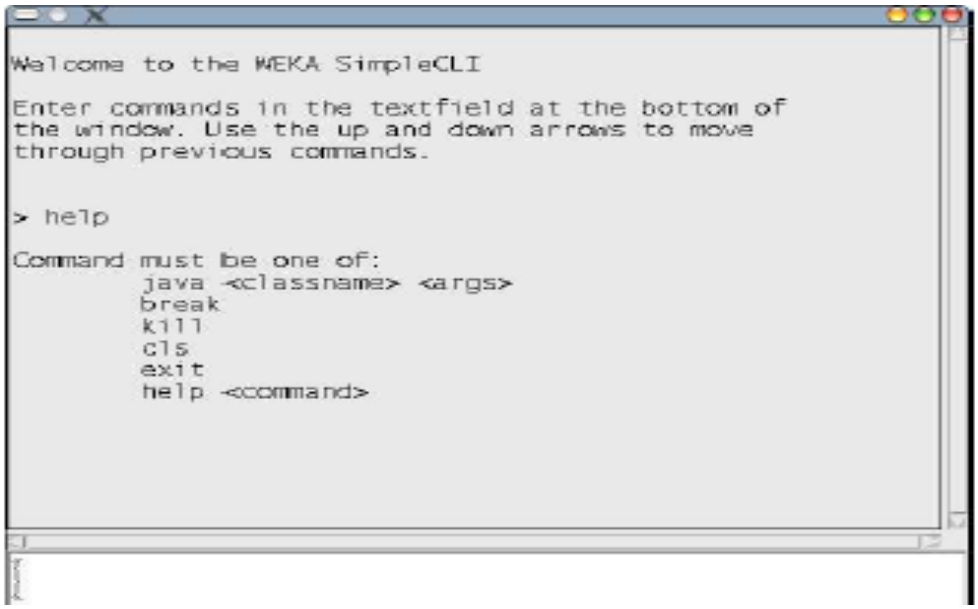


4.4 Simple CLI

Simple CLI es una abreviación de Simple Client. Esta interfaz proporciona una consola para poder introducir mandatos. A pesar de ser en apariencia muy simple es extremadamente potente porque permite realizar cualquier operación soportada por Weka de forma directa; no obstante, es muy complicada de manejar ya que es necesario un conocimiento completo de la aplicación.

Su utilidad es pequeña desde que se fue recubriendo Weka con interfaces. Actualmente ya prácticamente sólo es útil como una herramienta de ayuda a la fase de pruebas.

Figura 24: Interfaz de modo consola.



5. Stata

Stata es un software estadístico creado en 1985 por StataCorp de College Station, TX (800-STATA-PC). El nombre Stata es una abreviatura silábica de las palabras “Statistics” (Estadística) y “Data” (Datos).

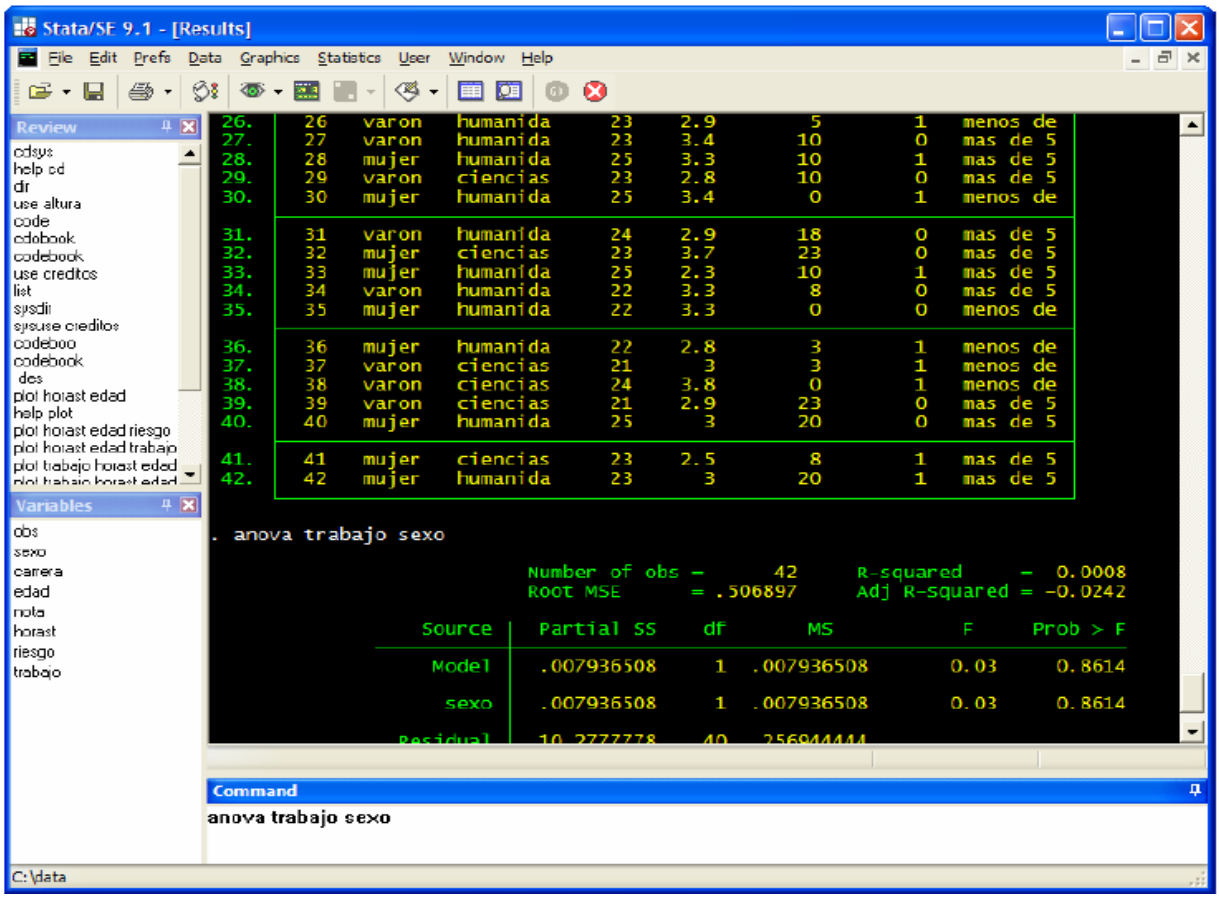
Esta herramienta proporciona una solución integrada de estadísticas, gráficos y gestión de datos para cualquier usuario que estudia datos, ya que está diseñado para el análisis descriptivo de datos y la implementación de diferentes técnicas de estimación, como métodos lineales, variables de tiempo, gráficos y métodos multivariables, entre otros, realizando análisis estadísticos sobre muestras aleatorias de poblaciones. Al contar con interfaces gráficas en un lenguaje sencillo, menús y cuadros de diálogos, este programa es utilizado en distintos campos de análisis e investigación por instituciones académicas y empresariales para estudios de economía, sociología, ciencias políticas, biomedicina y la epidemiología.

El software trabaja con grandes bases de datos que contienen información de diferentes variables para un conjunto de individuos o empresas. Distingue entre mayúsculas y minúsculas, de forma que por ejemplo las variables var1 y Var1 son distintas, lo que se debe tener en consideración al momento del desarrollo del trabajo. En cuanto a capacidad de manejar grandes volúmenes de información, a diferencia de otras aplicaciones, Stata necesita hacer una copia de la base de datos que se van a analizar en la memoria RAM del computador utilizado, por lo tanto, la memoria disponible en el computador deberá estar acorde con el tamaño de las bases de datos a utilizar.

5.1 Ambiente de trabajo Stata

La interface de Stata consiste en un entorno de trabajo que facilita la interacción con la aplicación, el cual tiene el siguiente aspecto:

Figura 25: Entorno de trabajo stata



La aplicación posee un lenguaje de programación avanzado que unas normas de sintaxis, al igual que otros lenguajes de programación como PASCAL o C++.

Cualquier orden en Stata (con muy pocas excepciones) posee la siguiente sintaxis:

Figura 26: Forma de ingreso comandos de Stata

command [varlist] [if] [in] [weight] [, options]

[…]	todo lo que aparece entre corchetes es opcional
if	seguida de una expresión lógica indica que sólo los datos que verifiquen dicha condición serán incluidos en el análisis
in	sirve para indicar el rango de observaciones que deseamos analizar
weight	sirve para indicar una variable de ponderación
options	son las opciones especificas del comando que estemos utilizando

Los ficheros de datos en Stata se denominan *dataset*. El cual es una tabla, donde las columnas representan variables y las filas de observaciones o casos.

Figura 27: Ejemplo de dataset.

sexo	carrera	edad	nota	horast	riesgo
mujer	ciencias	25	4	5	0
varon	humanida	28	3.3	5	1
mujer	humanida	25	3.3	0	1
mujer	humanida	24	2.2	20	0
varon	humanida	23	2.9	5	1
varon	humanida	23	3.4	13	0
mujer	humanida	25	3.3	10	1
varon	ciencias	23	2.8	10	0
mujer	humanida	25	3.4	0	1
varon	humanida	24	2.9	18	0
mujer	ciencias	23	3.7	20	0
mujer	humanida	25	2.3	10	1
varon	humanida	22	3.3	5	0
mujer	humanida	22	3.3	0	1
mujer	humanida	22	2.8	0	1
varon	ciencias	21	3	3	1
varon	ciencias	24	3.8	3	1
varon	ciencias	21	2.9	23	0
mujer	humanida	25	3	20	0
mujer	ciencias	23	2.5	5	0
mujer	humanida	23	3	20	0
mujer	ciencias	25	4	8	0
varon	humanida	28	3.3	8	1
mujer	humanida	25	3.3	3	1
mujer	humanida	24	2.2	23	0
varon	humanida	23	2.9	5	1
varon	humanida	23	3.4	10	0
mujer	humanida	25	3.3	10	1
varon	ciencias	23	2.8	10	0
mujer	humanida	25	3.4	0	1

5.2 Tipo de datos

La columna de storage type indica el formato de almacenamiento, es decir, el número de bytes y, por lo tanto, la precisión de la variable.

Figura 28: Ejemplo de tipos de datos mostrados en stata

```
. des

Contains data from C:\data\creditos.dta
  obs:            42
 vars:             8
 size:           630 (99.9% of memory free)

              storage  display     value
variable name   type   format      label      variable label

obs             byte   %8.0g
sexo            byte   %8.0g       sexo       sexo
carrera         byte   %8.0g       carrera    tipo de carrera cursada
edad            byte   %8.0g                  edad en años
nota            float  %9.0g                  nota sobre 5.
horast          byte   %8.0g                  horas trabajadas a la semana
riesgo          str1   %1s                    evaluacion del riesgo
trabajo         byte   %8.0g       trabajo

Sorted by:
```

5.2.1 Modelo de regresión por mínimos cuadrados

A diferencia de otras aplicaciones, en Stata los modelos de regresiones se ejecutan en dos fases claramente diferenciadas:

- Estimación de los parámetros del modelo.
- Diagnosis del modelo estimado.

El comando para realizar una estimación de los parámetros de un modelo de regresión lineal es:

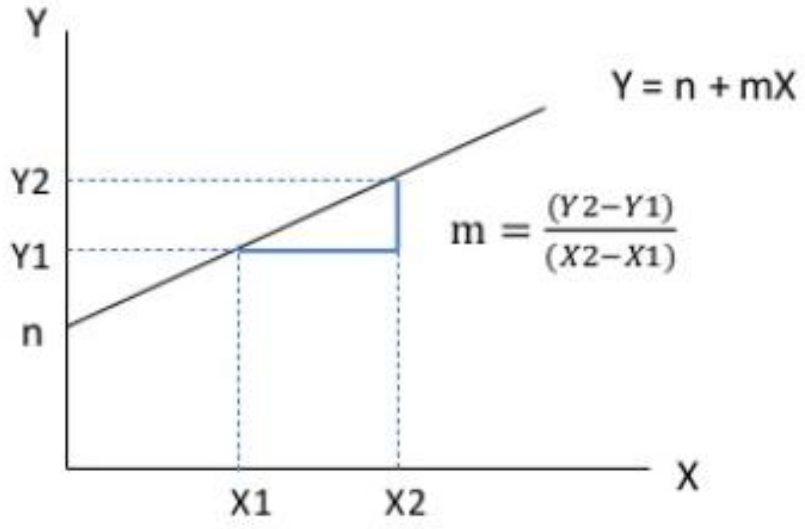
Figura 29: Comando de ingreso de regresión lineal.

```
regress depvar [indepvars] [if] [in] [weight] [, options]
```

5.2.2 Modelo de regresión lineal

La regresión lineal es una técnica estadística que permite determinar la relación existente entre una variable dependiente con una o más variables independientes. Permite describir y cuantificar las relaciones entre las variables y predecir valores para las variables independientes, y está basada en la ecuación de la recta:

Figura 30: Ecuación de la recta



La mejor forma de acercar esta recta, es ajustando los valores predichos y los valores observados. Por lo cual, la recta a elegir es aquella que minimice las distancias verticales entre las observaciones y el valor predicho. La forma utilizada para determinar la posición de la recta se denomina “Mínimo cuadrados ordinarios” y se hace reduciendo la siguiente función:

Figura 31: Fórmula para determinar la posición de la recta

$$\boldsymbol{Minimizar} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

En Stata se utiliza el comando *regress* y se visualiza de la siguiente manera:

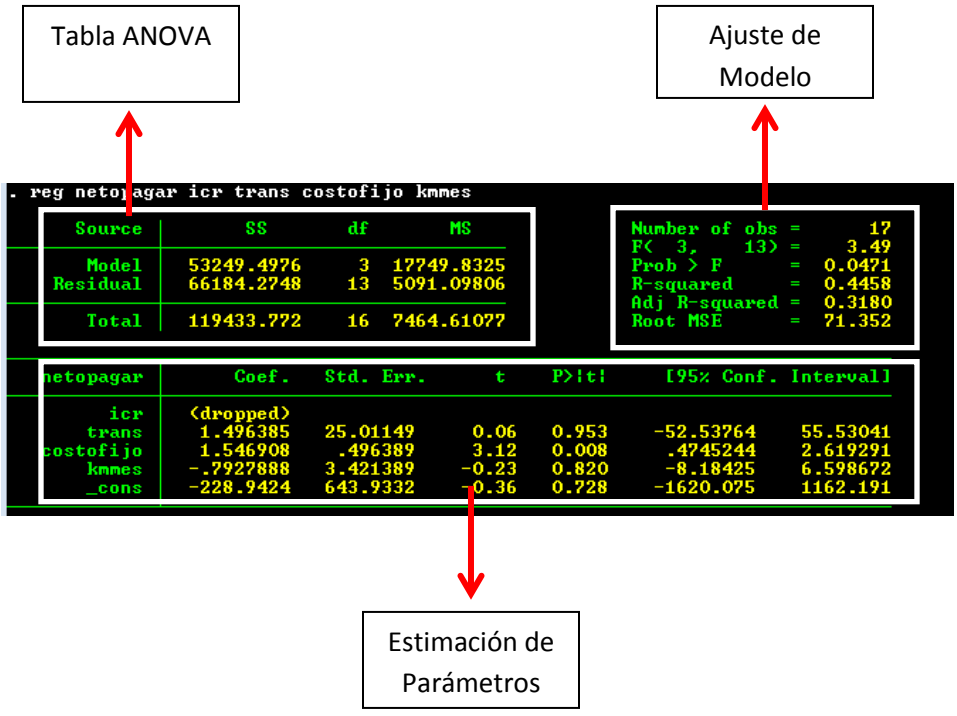
Figura 32: Regresión lineal en stata

```
Notes:
      1.  (/m# option or -set memory-) 1.00 MB allocated to data

. edit
(18 vars, 17 obs pasted into editor)

. reg netopagar icr trans costofijo kmmes

      Source |       SS       df       MS              Number of obs =      17
-------------+------------------------------           F(  3,    13) =    3.49
       Model |  53249.4976     3  17749.8325            Prob > F      =  0.0471
    Residual |  66184.2748    13  5091.09806            R-squared     =  0.4458
-------------+------------------------------           Adj R-squared =  0.3180
       Total |  119433.772    16  7464.61077            Root MSE      =  71.352

------------------------------------------------------------------------------
   netopagar |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         icr |  (dropped)
       trans |   1.496385   25.01149     0.06   0.953    -52.53764    55.53041
   costofijo |   1.546908    .496389     3.12   0.008     .4745244    2.619291
       kmmes |  -.7927888   3.421389    -0.23   0.820    -8.18425     6.598672
       _cons |  -228.9424   643.9332    -0.36   0.728    -1620.075    1162.191
------------------------------------------------------------------------------
```

5.3 Descripción de los componentes de la salida en Stata

Los componentes de salida se estructuran de la siguiente manera:

Figura 33: Salida de stata



5.3.1 Análisis de la varianza (anova)

a) *Source* (Fuentes): corresponde a las fuentes en que se descompone las varianza total del modelo para la variable dependiente. La varianza total se divide en la varianza que puede ser explicada por las variables independientes (Modelo) y la varianza no explicada, atribuible al término de error o residuo.

Figura 34: Source en stata

Source	SS	df	MS
Model	53249.4976	3	17749.8325
Residual	66184.2748	13	5091.09806
Total	119433.772	16	7464.61077

- **SS (SC):** son la suma de cuadrados de cada fuente que explica la varianza total.

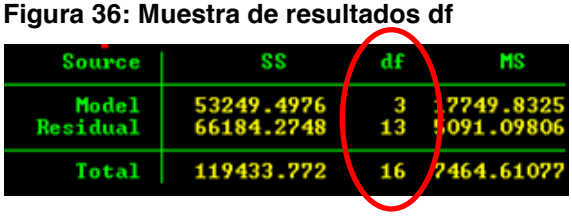
  SS-Total = SS-Modelo + SS-Residual

Figura 35: Muestra de resultados SS

Source	SS	df	MS
Model	53249.4976	3	17749.8325
Residual	66184.2748	13	5091.09806
Total	119433.772	16	7464.61077

- **df (gl):** son los grados de libertad. Para la varianza total los grados de libertad son N-1**.** Para el caso de la varianza del modelo los grados de libertad son iguales el número de parámetros estimados 2 (incluido intercepto) menos 1 (k-1). Finalmente los grados de libertad de la varianza del residuo se obtiene restando los *gl* totales menos los *gl* del modelo (N-1) – (k-1) = (N-k)-2. Donde “N” es el número de observaciones y “k” el número de parámetros estimados.

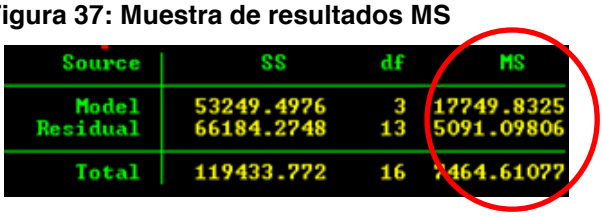
Figura 36: Muestra de resultados df



Source	SS	df	MS
Model	53249.4976	3	17749.8325
Residual	66184.2748	13	5091.09806
Total	119433.772	16	7464.61077

- **MS:** Es la desviación media, y se obtiene dividiendo la suma de los cuadrados por sus grados de libertad (SS/df).
    - MS-Modelo: SS-Modelo/(k-1)
    - MS-Residual: SS-Residual/(N-k-2)
    - MS-Total: SS-Total/(N-1)

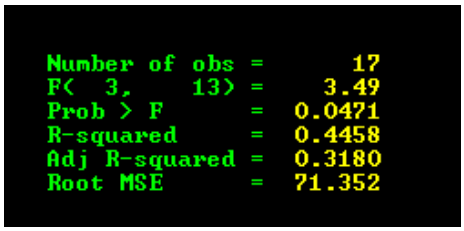
Figura 37: Muestra de resultados MS



Source	SS	df	MS
Model	53249.4976	3	17749.8325
Residual	66184.2748	13	5091.09806
Total	119433.772	16	7464.61077

b) Medidas de ajuste de modelo

Figura 38: Ajuste de modelo



```
Number of obs =       17
F(  3,    13) =     3.49
Prob > F      =   0.0471
R-squared     =   0.4458
Adj R-squared =   0.3180
Root MSE      =   71.352
```

F(3, 13): Este es el estadístico F de Fisher y corresponde al MS explicado dividido por el MS del residuo (MS-modelo/MS-Residual). En paréntesis aparecen los gl del ANOVA asociados a la varianza del modelo y la varianza del residuo. Este estadístico permite testear la significancia conjunta del modelo.

Cuando un modelo está muy mal especificado el modelo podría resultar no significativo.

- ***Prob > F*:** Este es el valor “p” asociado al estadístico F. Sirve para testear la hipótesis nula de que todos los parámetros del modelo (coeficientes) son iguales a cero.

- ***R-squared* (R cuadrado):** mide la bondad de ajuste del modelo, varía entre 0 y 1. Donde 0 es la ausencia de ajuste y 1 ajuste perfecto de la recta estimada. Se computa como la proporción de la varianza explicada por el modelo sobre la varianza total. La fórmula es la siguiente: (SS-Modelo/SS-Total). Este valor aumenta a medida que se incluyen más variables en el modelo.

- ***Adj R-squared* (R cuadrado ajustado):** Este estimado es un estimador del ajuste del modelo que penaliza la inclusión de nuevos regresores, es decir, no aumenta necesariamente e incluso podría disminuir. De este modo el R cuadrado ajustado busca dar una medida de la bondad de ajuste para obtener como resultado un modelo parsimonioso. La forma en que se computa este estadístico es como el cuociente entre la variación media del modelo y la variación media total: (MS-modelo/MS-Total).

- ***Root MSE* (Raíz del ECM):** es la raíz del error cuadrático medio, representa la desviación del término de error o residuo, y se obtiene como la raíz cuadrada del MS-Residual.

c) Estimación de Parámetros

Figura 39: Estimación de parámetros

netopagar	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
icr	(dropped)					
trans	1.496385	25.01149	0.06	0.953	-52.53764	55.53041
costofijo	1.546908	.496389	3.12	0.008	.4745244	2.619291
kmmes	-.7927888	3.421389	-0.23	0.820	-8.18425	6.598672
_cons	-228.9424	643.9332	-0.36	0.728	-1620.075	1162.191

En la primera fila aparecen las variables dependientes:

- ***Coef***: corresponde a los coeficientes estimados beta para la variable a analizar.
- ***Std. Err***.: corresponde al error estándar del coeficiente.
- ***t***: estadístico t para la hipótesis nula de coeficiente igual a cero.
- ***P>|t|***: es el p-value asociado al test.
- ***95% Conf. Interval***: intervalo de confianza al 95%

6. Resultados

En este capítulo se desarrollará el análisis completo de la metodología utilizada, se analizarán los resultados y se llegará a obtener la mejor propuesta para el modelo de comisiones, con el fin de justificar que es la opción más óptima al momento de compararla con la situación actual.

Figura 40: Diagrama de flujo de metodología de análisis de datos

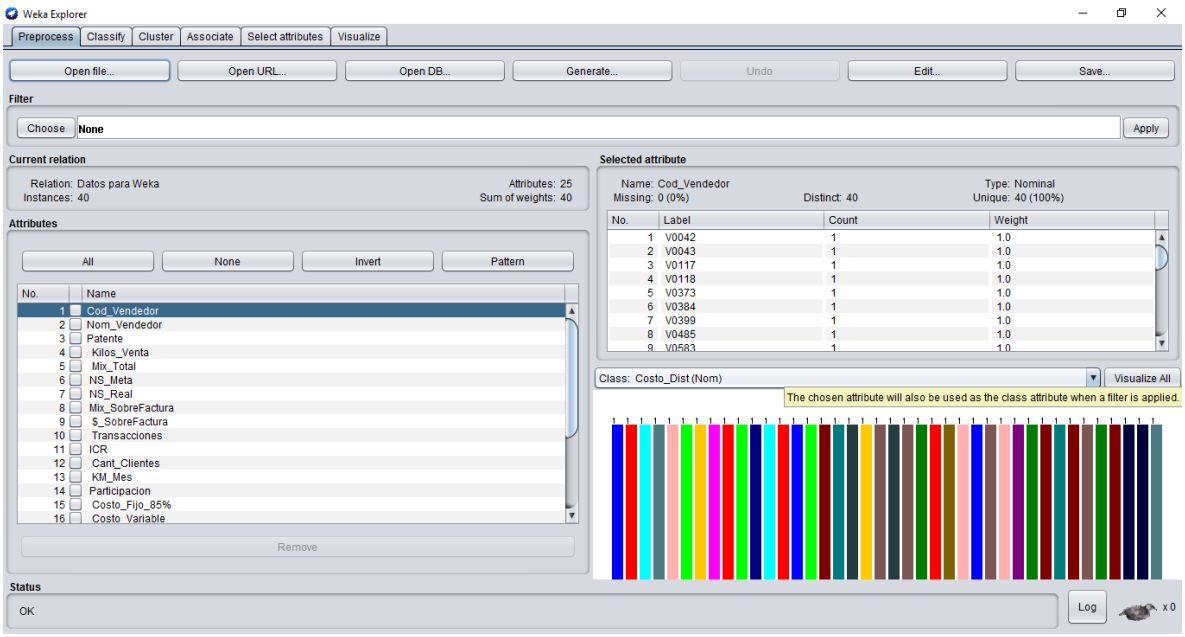
6.1 Análisis en Weka

Este programa se utiliza para analizar la base de datos y las variables utilizadas en el modelo de comisiones actual, para determinar si existe algún tipo de agrupación y que relevancia tienen estas variables.

6.1.1 Análisis en entorno explorer Preprocess

De acuerdo a lo explicado en capítulos anteriores, el entorno Preprocess se carga y manipulan los datos del modelo de comisiones de transportistas. Se realiza la carga sin realizar cambios en las variables ni suavizamiento de datos.

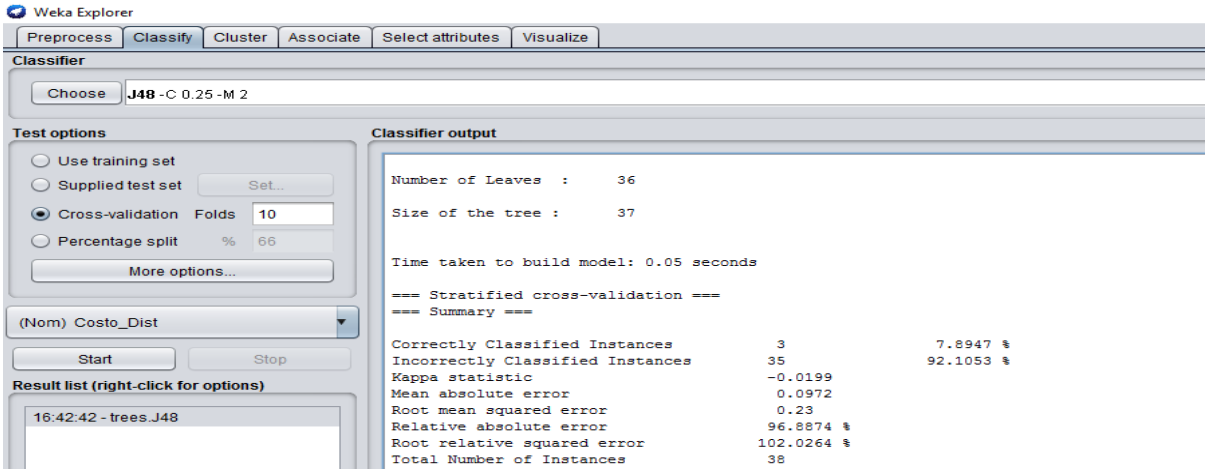
Figura 41: Resultado análisis de datos en explorer



6.1.2 Análisis en entorno explorer Classify

Dentro del entorno Classify se selecciona el modo *trees J48* (árbol de decisión), para someter a las diferentes variables de los datos originales y extraer relaciones generales que permitan generar hipótesis de comportamientos.

Figura 42: Resultado tree J48 en entorno Classify

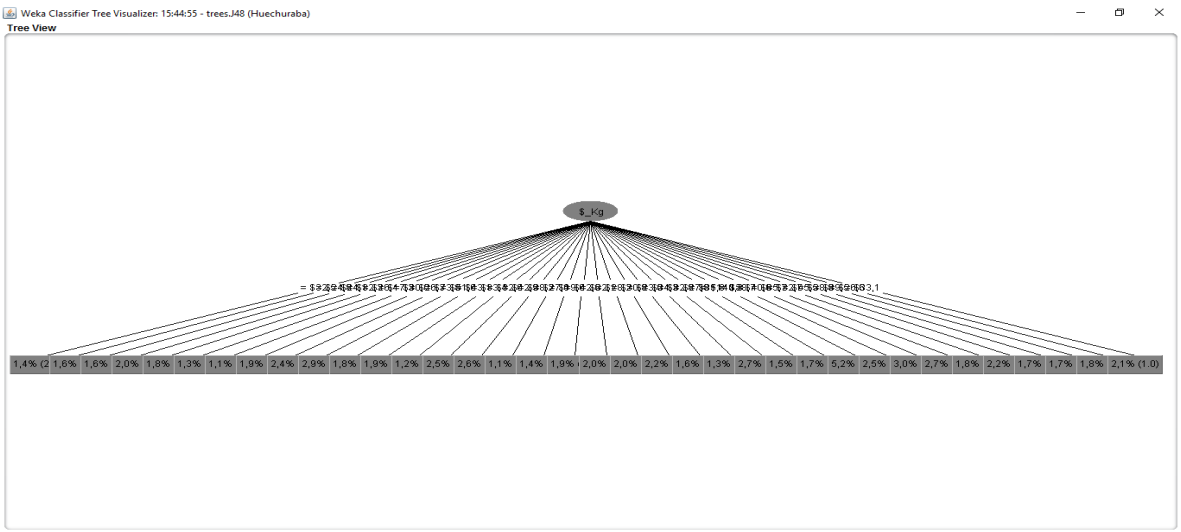


Los resultados obtenidos indican lo siguiente:

- Number of Leaves: 36 errores presentados de una muestra de 38.
- Correctly Classifies Instances: 3 instancias correctamente predichas.
- Incorrectly Classifies Instances: 35 instancias correctamente predichas.

En base a estos valores entregados por la metodología de árbol de decisión J48 se determina que no entrega y no cumple con lo requerido.

Figura 43: Visualización de tree J48

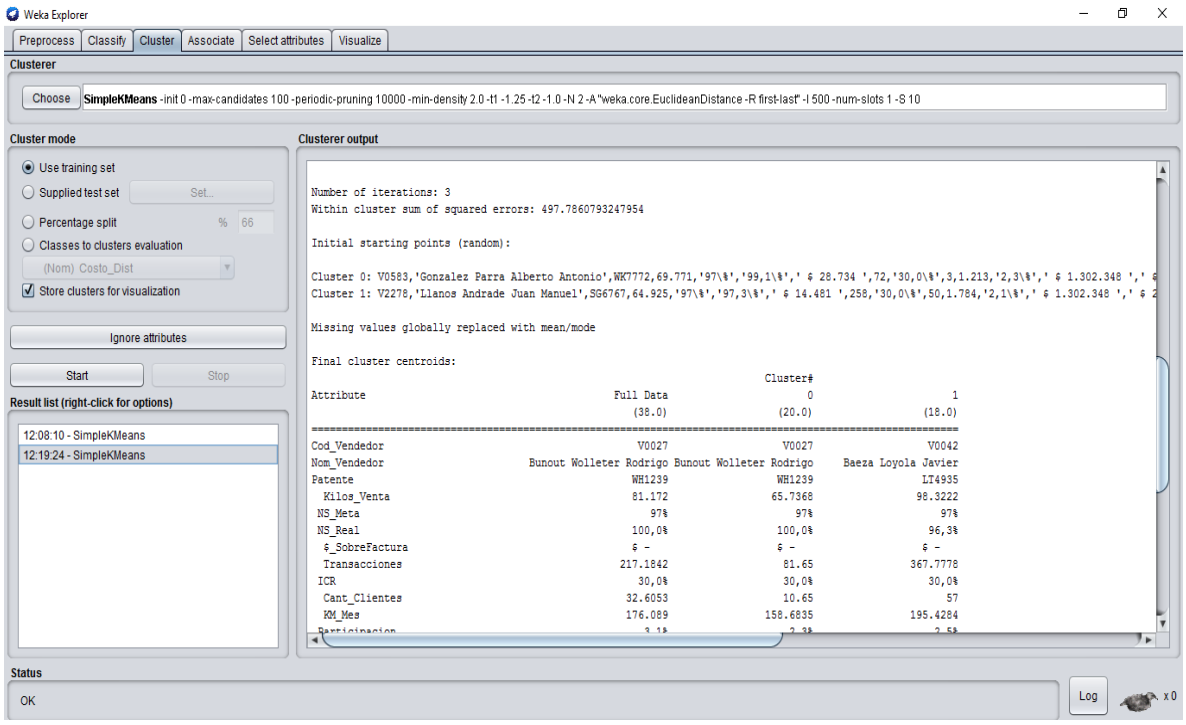


6.1.3 Análisis en entorno explorer Cluster

Este entorno permite aplicar algoritmos de agrupamientos de instancias de datos. Estos algoritmos buscan grupos de instancias con características “similares”, según un criterio de comparación entre valores de atributos de las instancias definidos en los algoritmos.

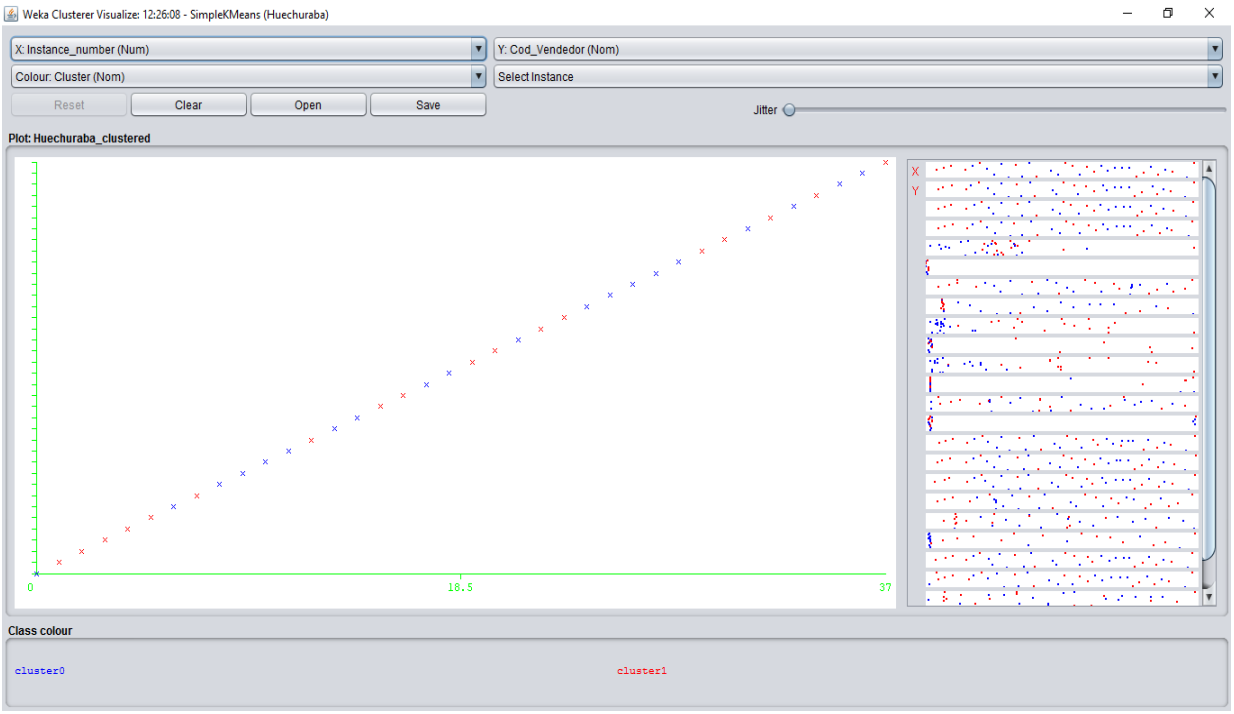
Se utiliza el algoritmo de agrupamiento *SimpleKmeans,* por ser uno de los más velices y eficientes, y precisa el número de categorías similares en las que se quiere dividir el conjunto de datos.

Figura 44: Resultado Cluster SimpleKmeans



Luego, se analiza gráficamente cómo se distribuyen los diferentes valores de los atributos en los grupos generados en la figura anterior obteniendo lo siguiente:

Figura 45: Visualización gráfica del agrupamiento de los atributos.



A la vista de esta gráfica se puede concluir que las variables que cuentan con características “similares” producto de su agrupamiento son:

- NS Meta
- Sobrefactura
- Transacciones
- ICR
- Kilómetros Mes
- Costo Fijo 85%
- Mermas

Para los análisis posteriores se decide dejar fuera la variable NS meta, ya que es la misma para todos los transportistas (97%).

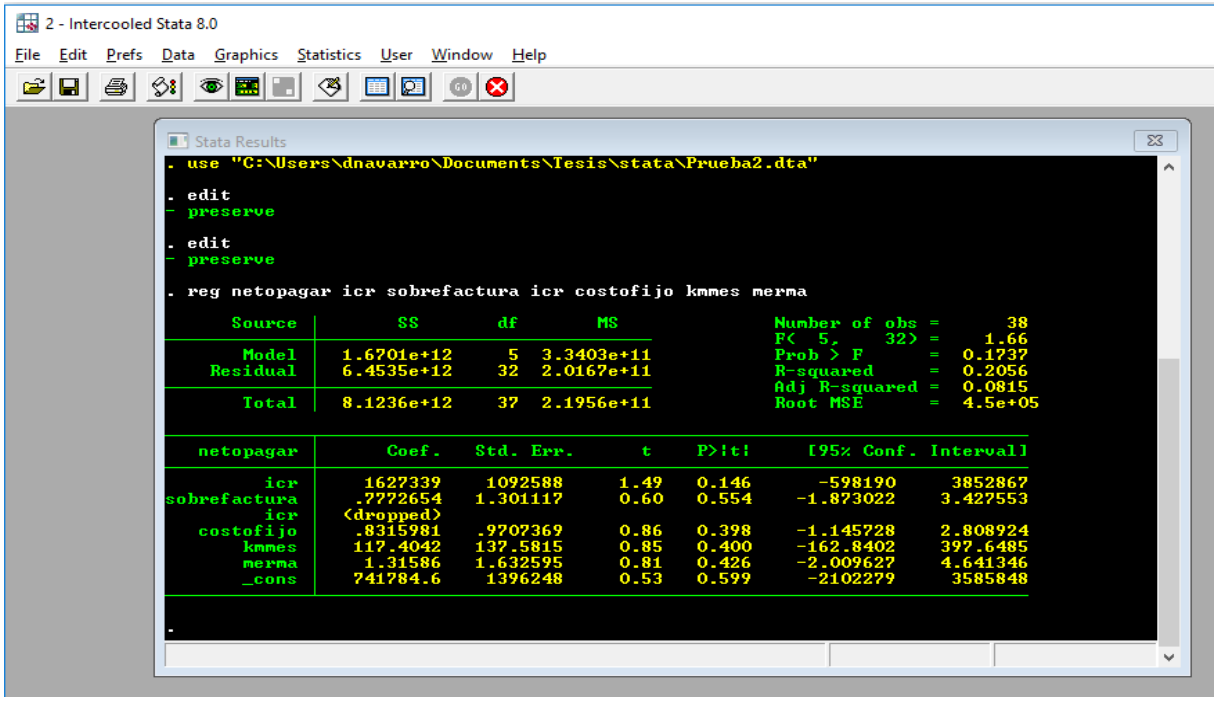
Con las variables ya definidas, se procede a trabajar en un análisis estadístico y técnicas de estimación.

6.2 Análisis en Stata

Con las variables a analizar ya definidas, tal como se indica anteriormente, se realiza la carga de datos en el programa Stata. El modelo utilizado es el de regresión lineal, lo cual permitirá determinar la relación existente entre una variable dependiente con una o más variables independientes.

Para este análisis primero se ingresan los datos puros y se obtienen los siguiente:

Figura 46: Primer análisis regresión lineal en Stata



Los resultados obtenidos indican lo siguiente:

- *R-squared*: 0,2056
- *Adj R-squared*: 0,0815

De acuerdo a estos resultados, un r2 de 20,56% indica una baja relación con una o más variables predictoras, por lo que se decide realizar el suavizamiento de datos.

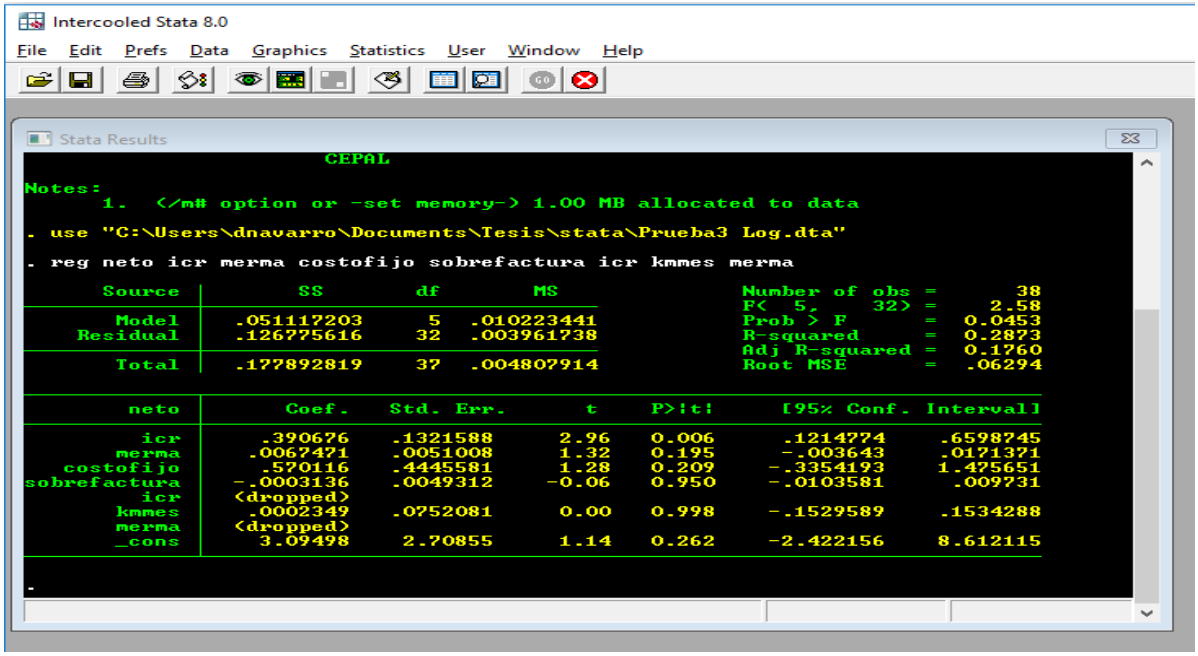
6.2.1 Suavizamiento de datos

Los datos a suavizar corresponde a todas las variables del modelo de comisión transportitas de la base de datos existente, para posteriormente ser caragada en Stata y determinar cuál metodología es la mejor a utilizar.

a) Logaritmo

Se aplica logaritmo a todas las variables del modelo de comisiones y se obtiene lo siguiente:

Figura 47: Resultados regresión con suavizamiento de dato con logaritmo.



Los resultados obtenidos son los siguientes:

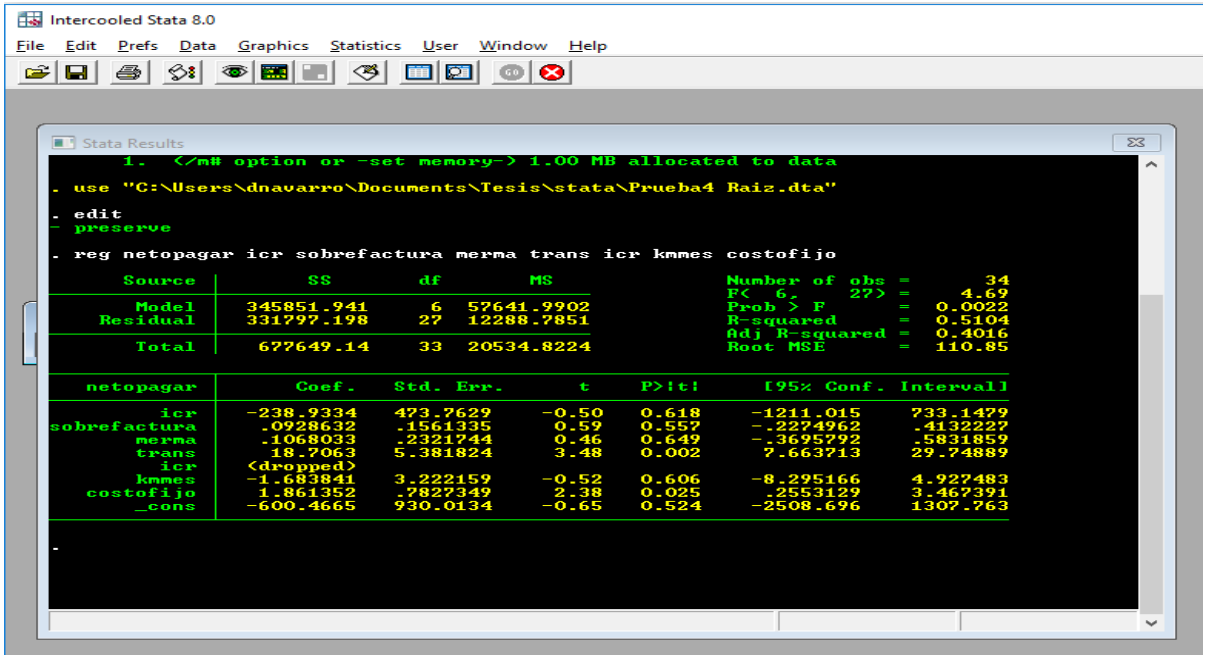
- *R-squared*: 0,2873
- *Adj R-squared*: 0,1760

Se observa un incremento en r2 y en el ajuste con respecto a los datos originales.

b) Raíz

Se aplica raíz a todas las variables del modelo de comisiones y se obvserva lo siguiente:

Figura 48: Resultados regresión con suavizamiento de dato con raíz.



Los resultados obtenidos son los siguientes:

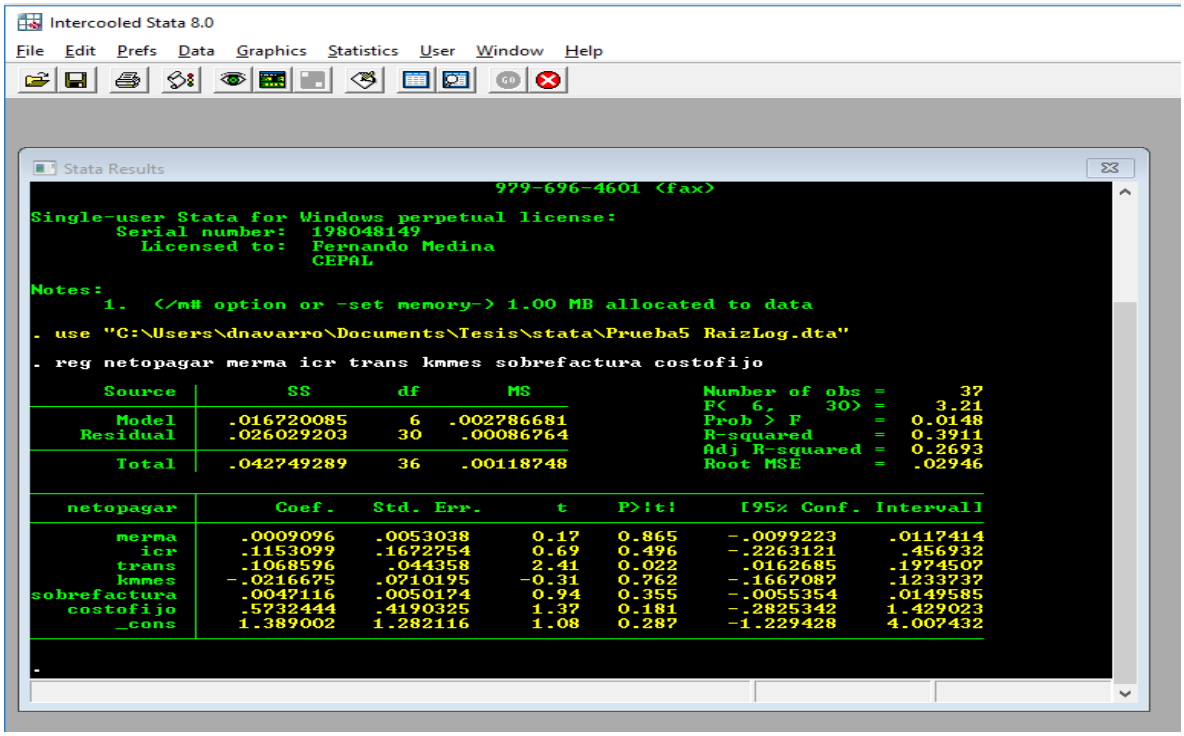
- *R-squared*: 0,5104
- *Adj R-squared*: 0,4016

Existe un incremento considerable de r2 y del ajuste con respecto a la alternativa anterior.

c) Logaritmo y raíz

Se aplica logaritmo y raíz a todas las variables del modelo de comisiones y se obvserva lo siguiente:

Figura 49: Resultados regresión con suavizamiento de dato con logaritmo y raíz.



Los resultados obtenidos son los siguientes:

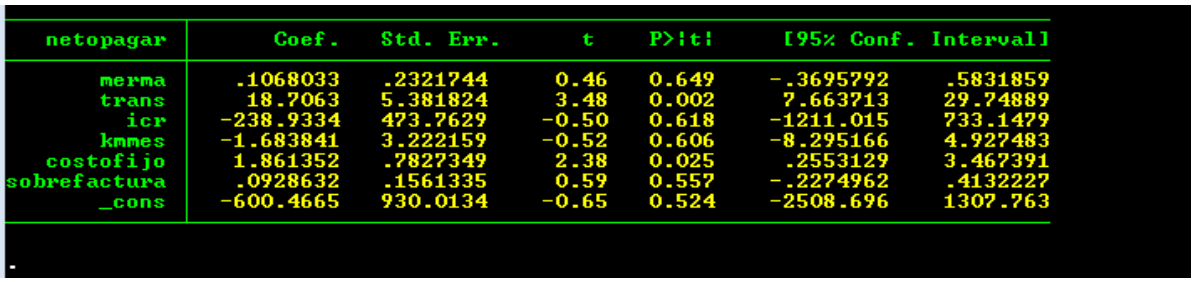
- *R-squared*: 0,3911
- *Adj R-squared*: 0,2693

Si bien los resultados obtenidos son mejores que la primera alternativa con logaritmo, no superan a lo obtenido con el suavizamiento de datos mediante la aplicación de raíz a las variables. Por lo tanto, la base de datos con la que se continuará trabajando será la de raíz.

6.2.2 Estimación parámetros del modelo

Con la base de datos ya definida, se realiza la carga en Stata para poder realizar el análisis de regresión lineal, el cual como ya ha sido mencionado anteriormente, permite describir y cuantificar las relaciones entre las variables y predecir valores para las variables independientes, para posteriormente realizar la estimación de parámetros y el análisis de los coeficientes que conforman la fórmula de pago de las comisiones:

Figura 50: Parámetros obtenidos en regresión lineal



netopagar	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
merma	.1068033	.2321744	0.46	0.649	-.3695792	.5831859
trans	18.7063	5.381824	3.48	0.002	7.663713	29.74889
icr	-238.9334	473.7629	-0.50	0.618	-1211.015	733.1479
kmmes	-1.683841	3.222159	-0.52	0.606	-8.295166	4.927483
costofijo	1.861352	.7827349	2.38	0.025	.2553129	3.467391
sobrefactura	.0928632	.1561335	0.59	0.557	-.2274962	.4132227
_cons	-600.4665	930.0134	-0.65	0.524	-2508.696	1307.763

.

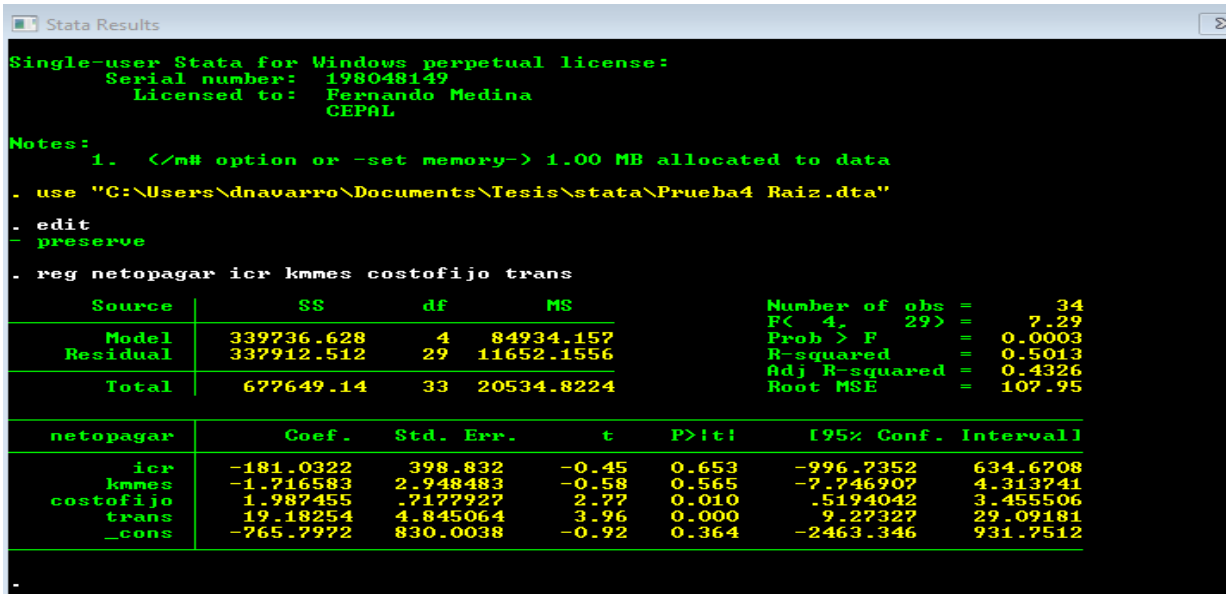
Neto a pagar = 0,1068 merma + 18,7063 transacciones – 238,9334 ICR – 1,6838 Km mes + 1,8614 costo fijo + 0,0929 sobrefctura.

Variables	Incidencia Neto a Pagar	Coeficiente
Índice complejidad de ruta	-	238,9334
Kilómetros mes	-	1,683841
Costos fijos	+	1,861352
Transacciones	+	18,7063
Sobrefactura	+	0,928632
Mermas	+	0,106833

Tabla 9: Incidencia de las variables sobre el neto a pagar de comisiones

Al analizar los coeficientes que confirman la fórmula de pago de las comisiones de transportistas se puede determinar que existen variables que tienen bajo peso o relevancia dentro del neto a pagar final, y estas son sobrefactura y mermas. La incidencia que tienen con respecto al neto final de comisiones es menor con al total acorde a su coeficiente: 0,0929 y 0,1068 respectivamente. En base a esto, se realiza nuevamente la regresión lineal para las variables ICR, Km mes, costos fijos y transacciones, obteniendo los siguientes resultados:

Figura 51: Segundo análisis regresión lineal



```
Stata Results
Single-user Stata for Windows perpetual license:
       Serial number:  198048149
         Licensed to:  Fernando Medina
                       CEPAL

Notes:
      1.  </m# option or -set memory-> 1.00 MB allocated to data

. use "C:\Users\dnavarro\Documents\Tesis\stata\Prueba4 Raiz.dta"

. edit
- preserve

. reg netopagar icr kmmes costofijo trans

      Source |       SS       df       MS              Number of obs =      34
-------------+------------------------------           F(  4,    29) =    7.29
       Model |  339736.628     4  84934.157             Prob > F      =  0.0003
    Residual |  337912.512    29  11652.1556            R-squared     =  0.5013
-------------+------------------------------           Adj R-squared =  0.4326
       Total |  677649.14     33  20534.8224            Root MSE      =  107.95

------------------------------------------------------------------------------
   netopagar |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         icr |  -181.0322   398.832    -0.45   0.653    -996.7352    634.6708
       kmmes |  -1.716583   2.948483   -0.58   0.565    -7.746907    4.313741
   costofijo |   1.987455   .7177927    2.77   0.010     .5194042    3.455506
       trans |   19.18254   4.845064    3.96   0.000     9.27327     29.09181
       _cons |  -765.7972   830.0038   -0.92   0.364    -2463.346    931.7512
------------------------------------------------------------------------------

.
```

De los datos obtenidos se desprende lo siguiente con respecto a las medidas de ajuste del modelo:

- *R-squared* y *Adj R-squared*: existe un incremento de ambos indicadores, 0,1102 y 0,1633 puntos respectivamente, con respecto a la regresión anterior con todas las variables consideradas.

Con respecto a los coeficientes se desprende lo siguiente:

Variables	Incidencia Neto a Pagar	Coeficiente
Índice complejidad de ruta	-	181,0322
Kilómetros mes	-	1,716583
Costos fijos	+	1,987455
Transacciones	+	19,18254

Tabla 10: Incidencia de las variables definidas

Comparando con los resultados de la primera regresión lineal, se observa que no existe una variación con respecto a la incidencia negativa o positiva de las variables en las comisiones de los transportistas.

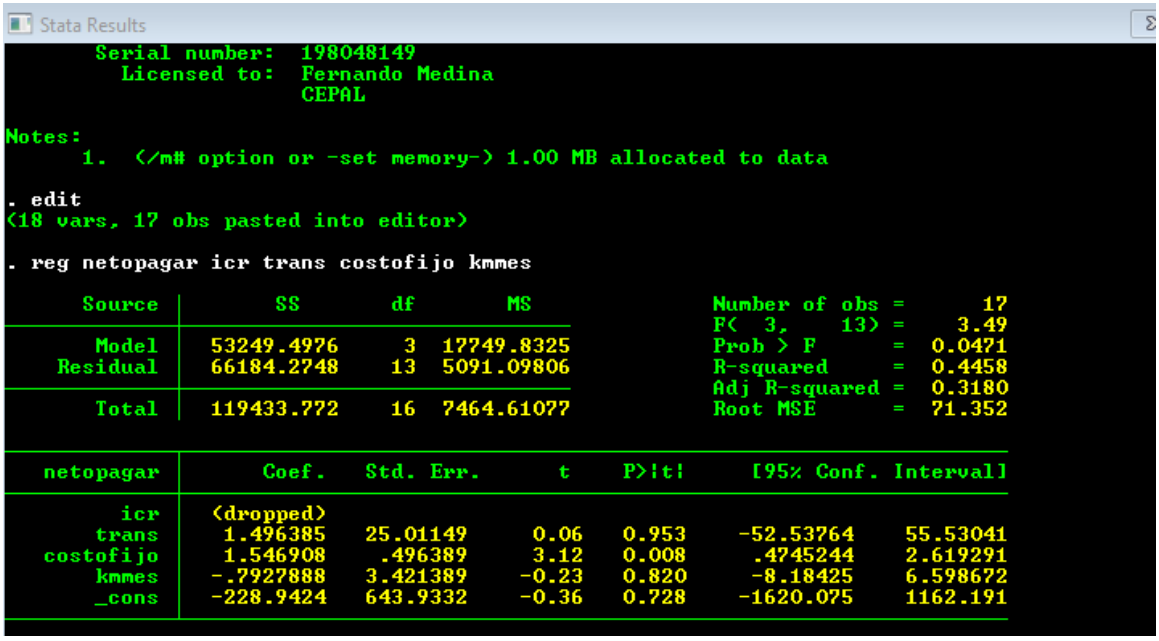
6.2.3 Simulación para dos tipos de rutas

Al revisar los datos obtenidos con las regresiones lineales anteriores, se decide simular en forma separada los dos casos que se dan dentro de las rutas que existen actualmente: atención exclusiva de supermercados y mixta (todo tipo de clientes). Para los casos prácticos se definen las siguientes siglas:

- **Ruta A:** entrega exclusiva a supermercados
- **Ruta B:** entrega a todo tipo de clientes.

Se realiza la regresión lineal para el caso de la ruta A obteniendo los siguientes resultados:

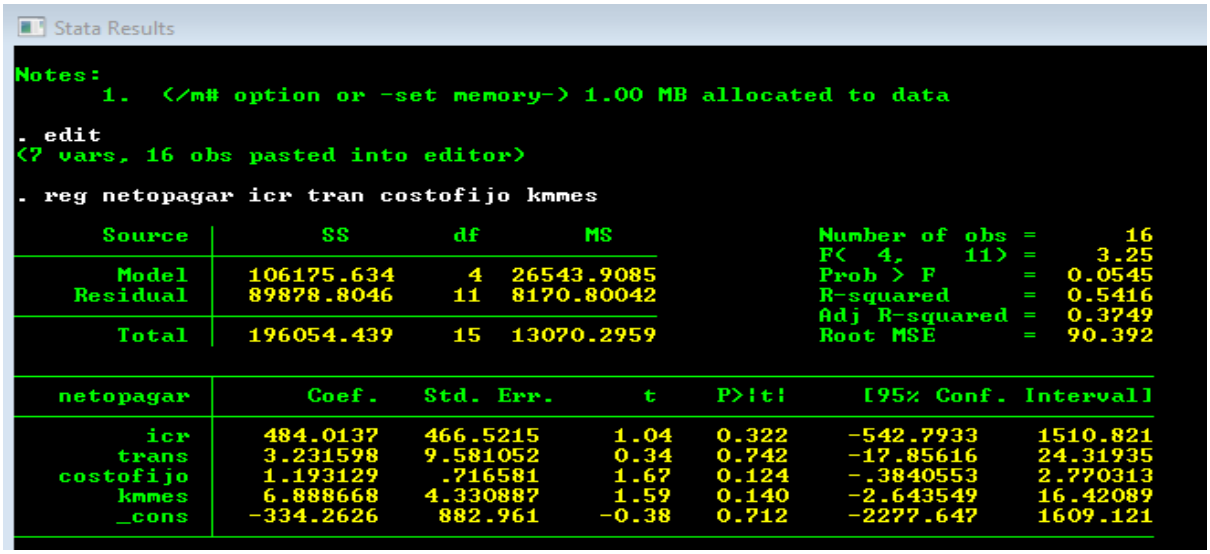
Figura 52: Resutaldo regresión lineal ruta A.



Neto Pagar = 1,5 transacciones + 1,55 costo fijo – 0,79 kilómetros mes – 228,94

Para la regresión de la ruta B se obtiene los siguientes resultados:

Figura 53: Resutaldo regresión lineal ruta B.



Stata Results

```
Notes:
      1.  (/m# option or -set memory-) 1.00 MB allocated to data

. edit
(7 vars, 16 obs pasted into editor)

. reg netopagar icr tran costofijo kmmes
```

Source	SS	df	MS		
Model	106175.634	4	26543.9085	Number of obs	= 16
Residual	89878.8046	11	8170.80042	F(4, 11)	= 3.25
				Prob > F	= 0.0545
				R-squared	= 0.5416
				Adj R-squared	= 0.3749
Total	196054.439	15	13070.2959	Root MSE	= 90.392

netopagar	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
icr	484.0137	466.5215	1.04	0.322	-542.7933	1510.821
trans	3.231598	9.581052	0.34	0.742	-17.85616	24.31935
costofijo	1.193129	.716581	1.67	0.124	-.3840553	2.770313
kmmes	6.888668	4.330887	1.59	0.140	-2.643549	16.42089
_cons	-334.2626	882.961	-0.38	0.712	-2277.647	1609.121

Neto Pagar = 484, 01 ICR + 3,23 transacciones + 1,19 costo fijo + 6,9 Km mes – 334,26

7.Conclusiones del análisis de resultados

Los programas Weka y Stata utilizados para el estudio de las comisiones de los transportistas actuales de Agrosuper, permitieron llegar a resultados concretos y validar su efectividad.
Los parámetros obtenidos dentro de las primeras simulaciones permiten concluir que:

- El Neto a Pagar a los transportistas disminuye cuando el ICR también baja, siendo esta la variable de mayor impacto en la comisión final, debido a que significa que existen menos visitas efectivas realizadas a clientes.
- Existe una leve baja de los kilómetros cuando el ICR disminuye, lo cual es explicado por la disminución de la densidad de clientes visitados. De forma contraria si existe un mayor tiempo de dedicación por cliente, aumenta el neto a pagar.
- En la medida que los costos fijos aumentan impactan linealmente el neto a pagar, pero sigue siendo ésta una variable de bajo impacto en las comisiones.
- El nivel de transacciones se mueve linealmente con el neto a pagar con un nivel de impacto mediano.
- La constante indica que el modelo parte desde un nivel negativo para el neto a pagar de los transportistas.

Para el caso de las simulaciones realizadas a dos tipos de rutas que existen dentro de los transportistas actualmente se concluye lo siguiente:

- **Ruta A:** en esta ruta específica para los supermercados, los coeficientes indican que al ser una ruta determinada previamente, recorrer una mayor cantidad de kilómetros significa una ineficiencia, y en los casos de los costos fijos y las transacciones su impacto es lineal con respecto al neto a pagar final, es decir, tienen una dependencia directa. Por lo tanto, en este caso no es productivo ni eficiente tanto para el transportista como para la empresa el salir de su ruta y visitar clientes que no hayan estado previamente asignados.

- **Ruta B:** en esta ruta determinada para todo tipo de clientes, se observa que los kilómetros son la variable con mayor peso dentro de las comisiones, porque bajo este tipo de ruta recorrer más kilómetros significa atender una mayor cantidad de clientes. Los costos fijos tienen un bajo peso con respecto a las transacciones, ya que de la forma en que está conformado el modelo de comisiones el monto a pagar de acuerdo al ICR puede llegar a ser mucho mayor incluso que el sueldo base. Incentivando a que estos pequeños distribuidores transporten productos independientes y adicionales ade la cantidad de clientes a atender en su ruta. Con esto, la empresa gana en tener mayores ventas de productos y menos pérdidas.

8. Bibliografía

1. [ALR] CN-2107-Applied-Logistic-regression.Download STATA. (disponible vía WEB en https://learn.canvas.net/courses/1179/pages/download-stata)

2. [BTCS09] Brain Trust Consulting Services. Técnicas para la Optimización de rutas de transporte y distribución. Septiembre 2009 (disponible vía WEB en http://www.odette.es/SGC/downloads/CAM/Vigilancia_Tecnologica_Tecnicas_Optimizacion_Rutas.pdf)

3. [MLGUW] Machine Learning Group at the University of Waikato. Downloading and installing Weka (disponible vía WEB en http://www.cs.waikato.ac.nz/ml/weka/downloading.html)

4. Tutorial WEKA 3.6.0. Ricardo Aler 2009. (disponible vía WEB en http://ocw.uc3m.es/ingenieria-informatica/herramientas-de-la-inteligencia-artificial/contenidos/transparencias/TutorialWeka.pdf)

5. [UCM] Universidad Carlos III de Madrid. Minería de Datos. (disponible vía WEB en http://www.it.uc3m.es/jvillena/irc/practicas/11-12/12mem.pdf)

6. [UISC12] Universidad ICESI. Diseño de Optimización del modelo de la red de distribución y transporte de empresa panificadora de productos de consumo masivo. Santiago de Cali 2012. (disponible vía WEB en https://repository.icesi.edu.co/biblioteca_digital/bitstream/10906/68156/1/dise%C3%B1o_optimizacion_modelo.pdf)

7. [UPV] Universidad Politécnica de Valencia. Un algoritmo para la optimización de rutas de transporte. Valencia España. (disponible vía WEB en http://users.dsic.upv.es/~agarridot/index_archivos/papers/garrido99b.pdf)

8. [UPC] Universidad Politécnica de Cartagena. Optimización con modelos de red en hoja de cálculo. (disponible vía WEB en http://www.uv.es/asepuma/XIII/comunica/comunica_17.pdf)