Importancia de las Analíticas Predictivas y dos populares herramientas que facilitan su uso, SPSS y SAS

Allan López Bastos
Universidad de Costa Rica,
Escuela de Ciencias de la Computación e Informática,
San José, Costa Rica
allan2786@gmail.com

Sergio Pastrana Espinoza
Universidad de Costa Rica,
Escuela de Ciencias de la Computación e Informática,
San José, Costa Rica
spastrana@gmail.com

Abstract

The predictive analysis is a tool that nowadays has become practically necessary for the good functioning of not only the big companies but also for the median and inclusive the small enterprise. Because of this fact, the development of statistical software packages has such a big importance since they facilitate in a significant way this labor. Being SPSS and SAS two of the most important software products.

Keywords: Predictive analytics, SPSS, SAS

Resumen

El análisis predictivo es una herramienta que hoy en día se ha vuelto prácticamente indispensable para el funcionamiento no solo de las grandes empresas sino también para la mediana e inclusive la pequeña empresa. Debido ha este hecho es que el desarrollo de paquetes de software estadísticos tiene una importancia tan grande ya que ellos facilitan de manera significativa esta labor. Siendo SPSS y SAS dos de los paquetes de software mas importantes que se encuentran en el mercado este artículo hará una pequeña introducción a ellos, sus principales características y funcionalidades; junto con una breve comparación entre ambos software.

Palabras clave: Analíticas predictivas, SPSS, SAS

1. Introducción

Hoy en día el posible éxito que pueden tener las empresas a futuro se ve influenciado por las decisiones que tomen sus dirigentes en el presente. Durante cada día en una empresa se generan cientos de decisiones críticas que afectarán la habilidad de la empresa de generar ganancias, manejo de riesgos, y alcanzar sus metas. Una forma de asegurar que se escojan muy buenas decisiones (talvez no las óptimas), es por medio de analíticas predictivas. Con ellas, las empresas analizan sus datos, combinan información de circunstancias pasadas, analizan sus presentes eventos, y deciden posibles futuras acciones a tomar, con lo que la mayoría de las empresas hoy en día se están convirtiendo en "Empresas Predictivas".
Muchas veces el resultado de aplicar estas analíticas predictivas es que se logran alcanzar exitosamente las metas organizacionales específicas, como pueden ser alcanzar cierto nivel de ganancias, reducir el costo de mercadeo o reducir comportamientos fraudulentos.
Las analíticas predictivas proveen una base cuantitativa para la rápida identificación, evaluación objetiva, y adquisición confidencial de nuevas oportunidades de mercado.
"Predictive analytics connects data to effective action by drawing reliable conclusions about current conditions and future events." [SPSS07].
A lo largo de éste artículo describiremos de forma general cómo funcionan las analíticas predictivas y algunas herramientas de software existentes para implementar dichos procesos, como son el SPSS y SAS (statistical analisys systems).

2. Analíticas predictivas

Las analíticas predictivas son un área del análisis estadístico que tiene que ver con la extracción de información a partir de datos, y usar esa información para predecir futuras amenazas y patrones de comportamiento. El núcleo de las analíticas predictivas se enfoca en la captura de relaciones entre variables preeditoras y las variables predecidas a partir de ocurrencias del pasado, para explotarlas de manera que se puedan predecir futuros eventos.
Las analíticas predictivas se componen de dos etapas principales para obtener los resultados óptimos

2.1 Funcionamiento de las analíticas predictivas:

Las analíticas predictivas se conforman de las siguientes dos fases:

- Analíticas avanzadas: análisis de circunstancias del pasado, presente y posibles circunstancias futuras, todo esto por medio de estadísticas, minería de datos, minería de textos, visualizaciones y reportes.
- Optimización de decisiones: para determinar si las acciones que se decidieron tomar son las derivarán en un mejor desempeño, y luego entregar las acciones óptimas al ente (ya sea una persona o un software) que pueda implementarlas de forma efectiva.

A parte existen tres tipos principales de analíticas predictivas cada una de las cuales tienen un propósito especifico de acuerdo con las necesidades que tenga la empresa y el tipo de información que se desea recabar.

2.2 Tipos de analíticas predictivas:

Entre las analíticas predictivas se encuentran las siguientes:

- **Modelos Predictivos**
  Los modelos predictivos analizan el desempeño pasado para averiguar la tendencia de ciertos clientes hacia ciertos comportamientos, de manera que en el futuro se puedan tomar en cuenta esas tendencias para mejorar la efectividad en el mercadeo. La mayoría de modelos predictivos realizan cálculos durante sus transacciones, como por ejemplo evaluar el riesgo de cierto cliente o transacción, para así llegar a una decisión.
- **Modelos Descriptivos**
  Los modelos descriptivos describen las relaciones en los datos de manera que se puedan clasificar los clientes o prospectos en grupos (identifican relaciones entre los clientes o los productos). Se utilizan, por ejemplo, para categorizar los clientes por sus preferencias respecto a los productos.

- **Modelos de Decisión**
  Describen la relación entre todos los elementos de una decisión (los datos conocidos, la decisión y el resultado esperado de la decisión) con el fin de predecir los resultados de decisiones que envuelven múltiples variables. Estos modelos se pueden utilizar en la optimización de decisiones.

Pasando a hablar de paquetes de software de este tipo se hará referencia a dos de los más importantes de nuestros días como lo son SPSS y SAS. El primero de estos se abordara a continuación haciendo una breve reseña historia y datos generales sobre sus características y su compatibilidad con el lenguaje de programación Python

3. Historia del SPSS

El programa SPSS (Statistical Product and Service Solutions) es un conjunto de potentes herramientas de tratamiento de datos y análisis estadísticos que funciona mediante menús desplegadles y cuadros de diálogo que permiten hacer la mayor parte del trabajo de manera sencilla.
Fue creado en 1968 por Norman H. Nie, C. Hadlai (Tex) Hull y Dale H. Bent. Originalmente el programa fue creado para grandes computadores. Y no es sino hasta en 1984 que sale la primera versión para computadores personales.

Todo esto llevo a que hoy en día el software posea grandes funcionabilidades y no solo se limite ha un paquete estadístico. Logrando atraer así a nuevos clientes y mejorando la utilidad que le pueden dar los antiguos clientes.

4. Generalidades de SPSS

Parte importante de la popularidad de SPSS se debe a la capacidad de trabajar con bases de datos de gran tamaño y permitir la recodificación de las variables y registros según como sean las necesidades del usuario.
SPSS analiza con detenimiento las variables implicadas en la investigación, con el propósito de construir un **modelo** único que sea capaz de explicar lo que aconteció, tanto antes como después del análisis estadístico. De alguna forma, SPSS trata de obtener información privilegiada a partir de la base de datos.
Otro aspecto de importancia es el hecho de que SPSS tiene muchas utilidades que son de gran utilidad para los usuarios que lo utilizan, ya que puede ser utilizado como:

- **Hoja de cálculo**
  SPSS permite realizar funciones aritméticas, algebraicas y trigonométricas sobre un fichero de datos. En este sentido, SPSS puede compararse, salvando las diferencias, a aplicaciones como Excel o Lotus.
- **Gestor de Bases de datos**
  SPSS permite gestionar de modo dinámico la información de un fichero de datos, pues se pueden actualizar los cambios operados (como ordenar, filtrar, etc.) o realizar informes personalizados de acuerdo con distintos criterios, etc. En este sentido, SPSS puede compararse, salvando las diferencias, a un gestor de bases de datos como Microsoft Access, Dbase, Oracle o Foxpro.
- **Generador de Informes**
  SPSS permite preparar de modo elegante atractivos informes de una investigación realizada, permitiendo incorporar en un mismo archivo – reporte el texto del reporte, las tablas y resultados estadísticos que el reporte necesite presentar e, incluso, los gráficos que se pudiesen generar. Todo ello apoyado por la posibilidad de exportar los reportes a una página web de modo completamente ágil. En este sentido, el paquete estadístico SPSS puede compararse, salvando las diferencias, a otros realizadores de reportes, como Microsoft Access.
- **Analizador de datos**
  SPSS tiene la capacidad de extraer de un fichero de datos toda la información recogida, ya sea superficial o profunda, permitiendo realizar procedimientos estadísticos descriptivos, inferenciales y multivariantes. En este sentido, SPSS puede compararse a programas como SAS, Statgraphics o Minitab.
- **Ejecutor de Minerías de Datos**
  SPSS puede llevar a cabo búsquedas inteligentes, para extraer información que permanecía oculta, elaborando árboles de decisión, segmentaciones de mercados o diseños de redes neuronales de inteligencia artificial. En este sentido, SPSS puede compararse a programas como SAS.

En tanto al aspecto grafico se refiere las el paquete cuenta con ambiente bastante amigable que permite que los nuevos usuarios se adapten pronto a la utilización del producto

5. Ventanas SPSS:

En cuanto a lo referente a la interfaz de SPSS cuenta con ocho tipos distintos de ventas. Cada una de ellas con una funcionalidad en particular entre las cuales podemos encontrar:

- El editor de datos: Contiene el archivo de datos sobre el que se basa la mayor parte de las acciones que es posible llevar a cabo con el SPSS. En esta ventana se pueden mostrar dos contenidos distintos, que son los datos propiamente y las variables del archivo y sus características que las definen.
- El visor de resultados: Recoge toda la información que el SPSS genera como consecuencia de las acciones que lleva a cabo. Además permite editar los resultados y guardarlos para su posterior uso.
- El editor de tablas: Permite editar los resultados que se presentan en formato de *tabla pivotante*
- El editor de gráficos: Permite modificar los colores, tipo de letra y demás detalles de los gráficos presentados en el visor de resultados.
- Editor de texto: Permite realizar cambios a los diferentes atributos (tipo, tamaño, color, etc) de los resultados de tipo texto.
- El borrador del Visor de resultados: presenta la misma información que el Visor de resultados pero únicamente en formato de texto, de manera que no permite realizar ningún tipo de edición a los datos obtenidos.
- El editor de sintaxis: Permite utilizar las opciones de programación de SPSS. Por medio de los botones Pegar disponible en la mayoría de los cuadros de diálogos permiten convertir las partes seleccionadas en editor de sintaxis SPSS para realizar modificaciones y guárdalas para luego utilizarlo en ejecuciones diferentes.
- Editor de procesos: Permite personalizar y automatizar algunas de las tareas que el SPSS lleva a cabo.

Además de todas estas facilidades el SPSS tiene la ventaja, como veremos en la siguiente sección, de permitir una integración con el lenguaje de programación Python y así incrementar la funcionalidad de ambas herramientas permitiendo la fusión de las mismas.

6. SPSS y Python

El Plug-In de Integración de SPSS y Python extiende la sintaxis de comandos del lenguaje de SPSS con las capacidades del lenguaje de programación Python. De ésta manera los programas Python pueden acceder ciertas propiedades de SPSS como son: información de diccionarios de variables, valores de salida de los procedimientos, y códigos de error de comandos SPSS. Se podrían utilizar estas características, por ejemplo, para crear dinámicamente una lista de variables SPSS del set de datos activos que tengan un atributo particular y después utilizar esa lista como la lista de variables para algún comando SPSS; o para realizar operaciones de administración de datos SPSS a un conjunto dinámico de archivos.

Se puede tener acceso total a todas las funcionalidades del lenguaje de programación Python por medio de los bloques entre los comandos BEGIN PROGRAM. – END PROGRAM. Por ejemplo un programa de “Hola, mundo!” sería:

```
BEGIN PROGRAM.
print “Hello, world!”
END PROGRAM.
```

Entre un mismo bloque entre los comandos anteriores, el procesador de Python tiene el control, por lo que todas las declaraciones deben ser declaraciones válidas de Python.

6.1. El módulo de Python *SPSS*

El módulo de Python *spss* viene instalado como parte del Plug-In de integración de SPSS-Python y contiene varias funciones específicas de SPSS que habilitan el proceso de usar la programación de Python con la sintaxis de comandos de SPSS. El módulo *spss* provee funciones para:

- Compilar y correr la sintaxis de comandos SPSS.
- Obtener información de datos de la actual sesión SPSS.
- Obtener datos, agregar nuevas variables y adjuntar casos al set de datos activo.
- Obtener resultados de salidas.
- Crear macro variables.

- Obtener información de errores.
- Manejar múltiples versiones del Plug-In de integración SPSS-Python.

Las funciones del módulo son acezadas incluyendo la declaración de Python *import spss* como la primera línea de un bloque de programa, por ejemplo:

```
BEGIN PROGRAM.
import spss
spss.Submit("SHOW ALL.")
END PROGRAM.
```

La declaración del import solo necesita ser incluida una vez durante una sesión SPSS. Incluir el comando en bloques de programas subsecuentes no tiene ningún efecto. El prefijo *spss* de la línea *spss.Submit* especifica que ésta función se encuentra en el módulo *spss*. Para funciones que serán utilizadas frecuentemente se puede incluir la línea *from spss import* <nombre de la función> antes de la primera llamada de la función, para no tener que incluirla cada vez que se va a utilizar la función, por ejemplo:

```
BEGIN PROGRAM.
import spss
from spss import Submit
Submit("SHOW ALL.")
END PROGRAM.
```

Como se verá en la siguiente sección, la función *Submit* nos permite mandar comandos a SPSS para ser procesados.

6.2. Someter comandos de SPSS

Es posible someter comandos de la sintaxis de SPSS entre bloques de programas Python por medio de la función *Submit* del módulo *spss*. La forma más simple de realizarlo es utilizando una cadena de caracteres entre comillas, representando un comando SPSS, y sometiendo el comando con la función antes mencionada. Por ejemplo:

```
BEGIN PROGRAM.
import spss
spss.Submit("FREQUENCIES VARIABLES=var1, var2, var3.")
END PROGRAM.
```

La sintaxis de comandos SPSS generada entre un bloque de programa y sometida a SPSS debe seguir ciertas reglas interactivas de sintaxis; por ejemplo, una cadena de comandos SPSS que se someta en un bloque de programa debe contener un punto (.) al final del comando. Además, el punto es opcional si el argumento de la función *Submit* contiene solo un comando.

7. Historia SAS

El SAS (Statistical Analysis System) es un sistema de software integrado que proporciona un control total sobre acceso, manejo, análisis y presentación de bases de datos.
El SAS fue creado en 1966 por Anthony J. Barr quien creó un análisis de varianza inspirado en la notación estadística de Maurice Kendall. En el año 1968 James Goodnight colaboro integrando regresiones múltiples y desarrollo rutinas de análisis de varianza. A partir de 1973 varias personas fueron colaborando con el desarrollo del proyecto hasta que en el 1976 se creo de manera formal el Instituto SAS.
Luego de todo este desarrollo se ha logrado crear un software de gran calidad en el ámbito de inteligencia analítica y de negocios de software y servicios; el cual cuenta con diversas características de gran interés para sus usuarios las cuales van en aumento con el pasar del tiempo.

8. Generalidades de SAS

Actualmente, los diferentes módulos de SAS hacen que éste sea un software de los que se llaman como "de inicio a fin". El cual permite entre otras cosas crear gráficos, trabajar como una hoja de cálculo, compilar programas en lenguaje C, incluye herramientas para construir interfases para la WWW, herramientas para tratar el Datawarehouse o para explotar datos con la filosofía del Datamining,
El ambiente de SAS básicamente es encuentra dividido en dos grandes ventanas:
- Ventana izquierda "Explorer": contiene accesos directos a los ficheros que interesen, información sobre las librerías y una ventana de resultados dónde aparece la información obtenida de las diferentes ejecuciones desglosadas.
- Ventana derecha: Contiene las ventanas principales **LOG**, **OUTPUT**, **EDITOR**.
El modo de trabajo que utiliza SAS se basa en éstas tres ventanas:
- Ventana EDITOR: Esta ventana corresponde a la ventana de sintaxis, por lo tanto es editable. Para poder ejecutar la sintaxis, se debe pulsar el botón  Para ejecutar una parte de la sintaxis, primero se selecciona dicha parte y después se pulsa el botón.
- Ventana LOG: En esta ventana se consulta y revisa todo lo que se ha ejecutado, aparecen mensajes de advertencia y de error en caso necesario y se informa sobre la velocidad de ejecución y recursos.
- Ventana OUTPUT: Cuando se ejecutan procedimientos de SAS, en esta ventana se muestran los listados, tablas y/o resultados.

9. SPSS versus SAS

Luego de conocer una poco sobre ambos software parece bastante conveniente realizar una pequeña comparación entre ambos para darnos una idea de sus principales diferencia y en que ocasiones deberíamos emplear SPSS y en cual SAS. Entre los aspectos más destacados para realizar esta comparación están:

- **Precio**
  El precio de SPSS es mucho menor que el del programa SAS, en torno a la mitad del precio. [3]
- **Rentabilidad**
  Recomendable para PYMES (hasta 500 trabajadores), SPSS es la mejor solución. Para grandes empresas (más de 500 empleados) o aquellas que puedan permitirse un fuerte desembolso sin necesidad de retorno a corto plazo, SAS es más rentable que SPSS, ya que permite ejecutar mayor número de procedimientos estadísticos y operativos. [3]
- **Facilidad**
  El manejo de SPSS es mucho más sencillo que el de SAS. El interfaz estilo hoja de cálculo de SPSS y su posibilidad de abrir ventanas muy comprensivas le convierten en un feo adversario para SAS. No obstante, SAS, una vez conocido el manejo de su lenguaje de programación, es más divertido que SPSS. [3]
- **Formación**
  La dependencia absoluta del lenguaje de programación por parte de SAS le hace muy vulnerable ante necesidades repentinas, ya que obliga a la Empresa a costear caros programas de formación, con el fin de permitir que su personal le saque el máximo provecho. SPSS, sin embargo, al ser mucho más fácil, no encadena al usuario a largos procesos formativos, sino que, en un tiempo mucho más corto que SAS, puede realizar complejos procedimientos de análisis sin esfuerzo. [3]
- **Robustez**
  La dependencia absoluta de Windows por parte de SPSS le hace muy vulnerable ante "caídas" del sistema, normalmente provocadas por acciones ilícitas del usuario. El programa SAS, por su parte, al tener su propio sistema operativo, toma la iniciativa del sistema, una vez arrancado, no permitiendo que una acción no autorizada por parte del usuario paralice el trabajo del operario. [3]

En la sección de anexos se pueden observar las principales pantallas que ofrecen éstas herramientas, lo que permite observar las similitudes y diferencias que existen entre ambos software.

10. Ejemplos

Para ilustrar un poco lo que es el trabajo utilizando este tipo de herramientas se incluyen dos pequeños ejemplos con su correspondiente explicación para darle al lector una pincelada de ambos software.

10.1. Ejempo en SPSS

```
*Abrimos el archivo
GET
FILE='C:\Archivos de programa\SPSS\Encuesta general USA 1991.sav'.
*Seleccionamos únicamente los casos correspondientes a mujeres.
USE ALL.
COMPUTE filter_$=(SEXO = 2).
VARIABLE LABEL filter_$ 'SEXO = 2 (FILTER)'.
VALUE LABELS filter_$ 0 'No seleccionado' 1 'Seleccionado'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .
*Guardamos únicamente los casos correspondientes a mujeres.
SAVE OUTFILE='C:\Archivos de programa\SPSS\Encuesta general USA 1991 reducida.sav'
/UNSELECTED = DELETE
/COMPRESSED.
*Abrimos el nuevo archivo para comprobar que sólo contiene los registros correspondientes a *mujeres.
GET
FILE='C:\Archivos de programa\SPSS\Encuesta general USA 1991 reducida.sav'.
```

El código de ejemplo anterior abre un archivo por medio del comando GET FILE =´ ´.
Después busca en el archivo los casos que corresponden a mujeres filtrando la información del archivo por medio de varias instrucciones. Luego con el comando SAVE OUTFILE= ´ ´. guarda en otro archivo nuevo solamente los casos antes filtrados pero sin almacenar también el filtro, esto por medio de los comandos /UNSELECTED = DELETE y /COMPRESSED. Al final se abre el nuevo archivo creado con los datos de solo las mujeres sin ningún campo extra.

10.2. Ejemplo en SAS

Ejemplo con su respectiva explicación obtenida de [5]

```
DATA PACTIVO2;
INFILE 'A:\DATOS.DAT' LRECL=9;
INPUT NUM_PAC 1-2 INIC $ 4-6 SEXO 7 EDAD 8-9;
RUN;
PROC PRINT DATA=PACTIVO2;
RUN;
DATA PACTIVO3;
INFILE 'A:\DATOS2.DAT' DLM='09'x;
INPUT NUM_PAC TRATAM;
RUN;
PROC PRINT DATA=PACTIVO3;
RUN;
```

La instrucción INFILE se utiliza para la lectura de datos externos y en ella se menciona la ruta dónde es encuentra el fichero que contiene los datos. La opción LRECL de la instrucción INFILE indica la longitud máxima de cada línea (es indispensable si cada registro tiene más de 256 caracteres).
En la instrucción INPUT se declara las variables que se van a leer. Se escriben las columnas dónde se encuentran las variables si el fichero de datos externo es de formato fijo. En el caso en qué el fichero de datos está delimitado, no tiene sentido especificar las columnas. Por defecto, el separador que lee SAS ® es el espacio, pero con la opción DLM se

define el delimitador que deseado, por ejemplo: DLM='09'x si el fichero es encuentra delimitado por tabuladores o DLM=';' si el fichero es encuentra delimitado por el símbolo " ; ".
Cualquier procedimiento trabaja con el dataset deseado utilizando la opción DATA=nombre_dataset. Por defecto, SAS utiliza el dataset creado en el paso DATA más reciente.
Es útil observar el listado producido por PROC PRINT para comprobar que efectivamente los datos se han leído perfectamente (en ocasiones no se cometen errores en la sintaxis aunque los datos no se leen adecuadamente).

11. Conclusiones

A pesar de que SPSS parece tener mas ventajas respecto a SAS, no podemos generalizar diciendo que es mejor, son solo dos opciones distintas.

El usar este tipo de herramientas en una empresa, si bien no es obligatorio, es un factor que puede influir de manera directa en el desempeño de la misma.

Las herramientas de éste tipo son bastante útiles no solo por permitirnos realizar analíticas predictivas, sino también por ser programas de muchas aplicaciones en una.

12. Referencias

[1] www.uca.es/serv/ai/formacion/spss/Inicio.pdf

[2] http://es.wikipedia.org/wiki/SPSS

[3] http://www.estadistico.com/arts.html?20001113

[4] http://www.unalmed.edu.co/~estadist/Esta1/INDUCCION%20SAS.pdf

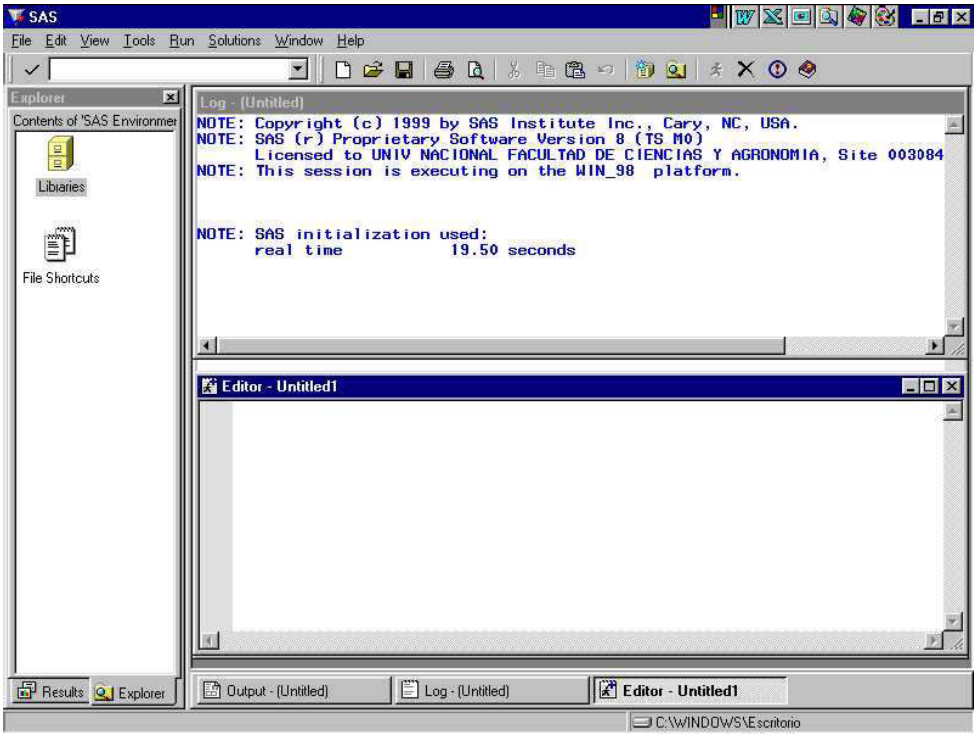
[5] http://einstein.uab.es/_c_serv_estadistica/Manuals/ManualSAS.PDF

[6] http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Progra/SAS_V8_V1_2.pdf

[7] http://estadistica.ieg.csic.es/tutoriales/PDF/SintaxisSPSS.pdf

12. Anexos

12.1. Ventana de SAS



12.2. Ventana SPSS