GESTIÓN GLOBAL DEL RIESGO. SCORING

CAIO FERNANDES MORENO

UCM - Mineria de Datos

TAREA 3 Credit Scoring (segunda tarea del profe Escot)

# 1. Enunciado del trabajo

Esta es la segunda práctica que hay que entregar al profesor Escot. Se trata de construir una tarjeta de puntuación del riesgo de crédito para clientes de tarjeta de crédito. Y hacer previsiones sobre nuevos clientes ¿le daríais la tarjeta a esos nuevos clientes?.

En el Excel hay información sobre clientes a los que se le ha dado la tarjeta de crédito (Cardhldr=1) y sobre los que sabemos si han impagado alguna vez o no (default= 1 o 0 respectivamente). También hay clientes rechazados (Cardhldr=0) que son clientes a los que no se les concedió la tarjeta, y por tanto no sabemos si hubieran impagado o no (default= na).

Por último al final del archivo hay 34 nuevos clientes (Cardhldr=na) individuos con identificador de cliente desde 1286 hasta 1319, que solicitan una tarjeta de crédito.

Tenéis que construir el modelo de Scoring siguiendo la metodología SAS Miner. Si queréis construir algún otro modelo (redes neuronales, árboles, o randomforest) también podéis hacerlo. En cualquier caso tenéis que seguir la metodología expuesta en clase, y en las presentaciones de Javier Monjas y Caridad Pavón (que esperemos que pueda venir en enero)

El resto de variables son las siguientes:

Age = Age n years plus twelfths of a year

Income = Yearly income (divided by 10,000)

Exp\_Inc = Ratio of monthly credit card expenditure to yearly income

Avgexp = Average monthly credit card expenditure

Ownrent = 1 if owns their home, 0 if rent

Selfempl = 1 if self employed, 0 if not.

Depndt = 1 + number of dependents

Inc\_per = Income divided by number of dependents

Cur\_add = months living at current address

Major = number of major credit cards held

Active = number of active credit accounts

Os recuerdo las Fases

FASES DEL ANALISIS

1) Delimitación del estudio: objetivo, limitaciones y disponibilidad de datos

2) Análisis descriptivo exploratorio de los datos

¿Qué variables tenemos? ¿Cuál es la variable objetivo?

¿Qué porcentaje de buenos y malos tenemos en cada categoría?

¿Hay datos anómalos?

¿Hay datos perdidos?

¿Hay que imputar?

3) ¿Hay que tramificar variables?

4) Selección de variables ¿qué variables son las más importantes?¿qué variables deben incluirse en el scorecard? disponibilidad, e importancia, estadísticos de selección (IV, correlación)

5) Muestreo:

Muestra de desarrollo/entrenamiento (70-80%) y de validación (30-20%)

Sobre muestreo de buenos/malos?

6) Estimación del primer modelo, diagnosis (Kolmogorov-Smirnov, c-statistic, Gini) y obtención del scorecard preliminar

7) Inferencia de denegados

8) Estimación del modelo definitivo y obtención del scorecard definitivo

9) Validación y Seguimiento

# 2. Ejecución del trabajo

Para el trabajo he utilizado en fichero Excel llamado **datospracticas\_alumnos\_OK.xls.**

Lo primero que hay que hacer es abrir el fichero **datospracticas\_alumnos\_OK.xls** y guardar el archivo en formato Excel 95 (**datospracticas\_alumnos\_OK\_excel95.xls)** para poder importar los datos con SAS Base.

Manualmente he creado 3 archivos Excel:

**CS\_REJECTS.xls** = Todos los clientes con Cardhldr=0 y default vacío, o sea clientes con tarjeta que ha sido denegada.

Contiene 291 observaciones.

**CS\_NEW\_CLIENTS.xls** = Todos los clientes que tienen default y Cardhldr vacío.

Contiene 34 observaciones.

**CS\_ACCEPTS.xls** = Todos los clientes con Cardhldr=1 (aceptados para tener tarjeta de crédito) y default que no sea vacío.

Contiene 994 observaciones.

A bajo el código SAS para importar los datos y crear los datasets.

**PROC** **IMPORT** OUT= riesgo.cs\_all

DATAFILE= "C:\Users\win\Documents\GitHub\ucm\score\trabajo3\datospracticas\_alumnos\_OK\_excel95.xls"

DBMS=EXCEL5 REPLACE;

GETNAMES=YES;

**RUN**;

**PROC** **IMPORT** OUT= riesgo.cs\_all\_without\_new\_clients

DATAFILE= "C:\Users\win\Documents\GitHub\ucm\score\trabajo3\datospracticas\_alumnos\_OK\_excel95\_accepts\_rejects\_without\_new\_clients.xls"

DBMS=EXCEL5 REPLACE;

GETNAMES=YES;

**RUN**;

**PROC** **IMPORT** OUT= riesgo.cs\_rejects

DATAFILE= "C:\Users\win\Documents\GitHub\ucm\score\trabajo3\datospracticas\_alumnos\_OK\_excel95\_rejects.xls"

DBMS=EXCEL5 REPLACE;

GETNAMES=YES;

**RUN**;

**PROC** **IMPORT** OUT= riesgo.cs\_accepts

DATAFILE= "C:\Users\win\Documents\GitHub\ucm\score\trabajo3\datospracticas\_alumnos\_OK\_excel95\_accepts.xls"

DBMS=EXCEL5 REPLACE;

GETNAMES=YES;

**RUN**;

**PROC** **IMPORT** OUT= riesgo.cs\_new\_clients

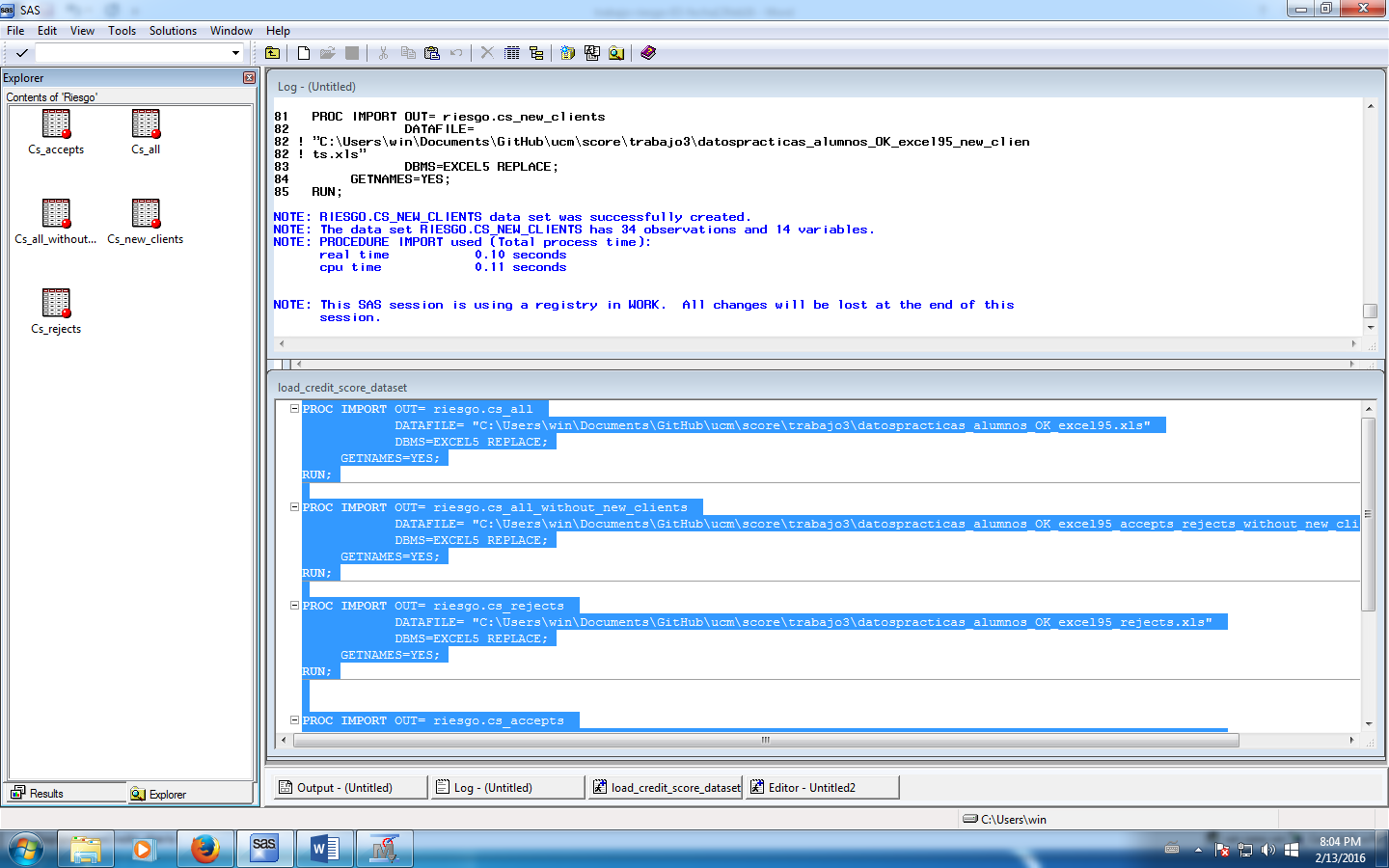
DATAFILE= "C:\Users\win\Documents\GitHub\ucm\score\trabajo3\datospracticas\_alumnos\_OK\_excel95\_new\_clients.xls"

DBMS=EXCEL5 REPLACE;

GETNAMES=YES;

**RUN**;

Abajo la imagen de los 5 datasets creados.



1. **Delimitación del estudio: objetivo, limitaciones y disponibilidad de datos**

El objetivo del trabajo es construir el modelo de Scoring siguiendo la metodología SAS Miner, con tarjeta de puntuación del riesgo de crédito para clientes de tarjeta de crédito es posible predecir sobre los 34 nuevos clientes si se debe o no dar crédito.

Abajo hay informaciones importantes dadas en el enunciado del trabajo:

“*En el Excel hay información sobre clientes a los que se le ha dado la tarjeta de crédito (Cardhldr=1) y sobre los que sabemos si han impagado alguna vez o no (default= 1 o 0 respectivamente). También hay clientes rechazados (Cardhldr=0) que son clientes a los que no se les concedió la tarjeta, y por tanto no sabemos si hubieran impagado o no (default= na).*

*Por último al final del archivo hay 34 nuevos clientes (Cardhldr=na) individuos con identificador de cliente desde 1286 hasta 1319, que solicitan una tarjeta de crédito.”*

Notas:

Cardhldr = 1 = Ha recebido la tarjeta

Cardhldr = 0 = Clientes rechazados, no han recibido la tarjeta de crédito, tienen el default = na.

34 nuevos clientes con cardhldr = na de la línea 1286 hasta 1319 que solicitaron una tarjeta de crédito.

1. **Análisis descriptivo exploratorio de los datos**

**¿Qué variables tenemos?**

Hay un total de 14 variables existentes en el dataset y 1319 líneas (rows).

Son ellas:

* ID = Identificador
* Cardhldr = 1 ha recebido la tarjeta / 0 clientes rechazados
* Default = 1 no ha pagado (malos) / 0 ha pagado (buenos)
* Age = Age n years plus twelfths of a year
* Income = Yearly income (divided by 10,000)
* Exp\_Inc = Ratio of monthly credit card expenditure to yearly income
* Avgexp = Average monthly credit card expenditure
* Ownrent = 1 if owns their home, 0 if rent
* Selfempl = 1 if self employed, 0 if not.
* Depndt = 1 + number of dependents
* Inc\_per = Income divided by number of dependents
* Cur\_add = months living at current address
* Major = number of major credit cards held
* Active = number of active credit accounts

Abajo una tabla con las variables y el tipo de cada variable definido por mi.

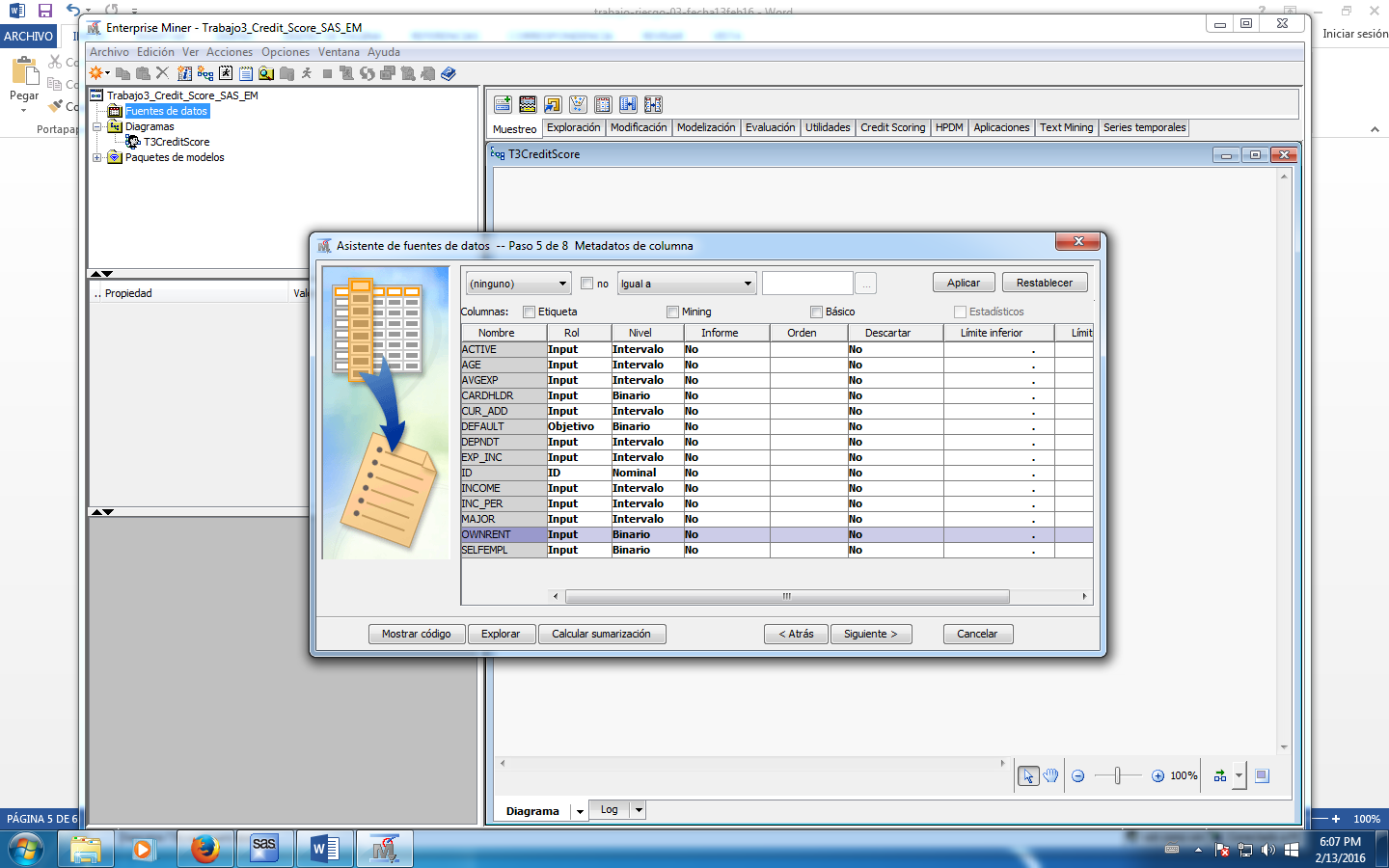
|  |  |  |
| --- | --- | --- |
| **Variable** | **Tipo** | **Rol** |
| ID | Nominal | ID |
| Cardhldr | Binaria | Input |
| Default | Binaria | Objetivo |
| Age | Intervalo | Input |
| Income | Intervalo | Input |
| Exp\_Inc | Intervalo | Input |
| Avgexp | Intervalo | Input |
| Ownrent | Binaria | Input |
| Selfempl | Binaria | Input |
| Depndt | Intervalo | Input |
| Inc\_per | Intervalo | Input |
| Cur\_add | Intervalo | Input |
| Major | Intervalo | Input |
| Active | Invervalo | Input |

**¿Cuál es la variable objetivo?**

La variable objetivo es la llamada **Default.**

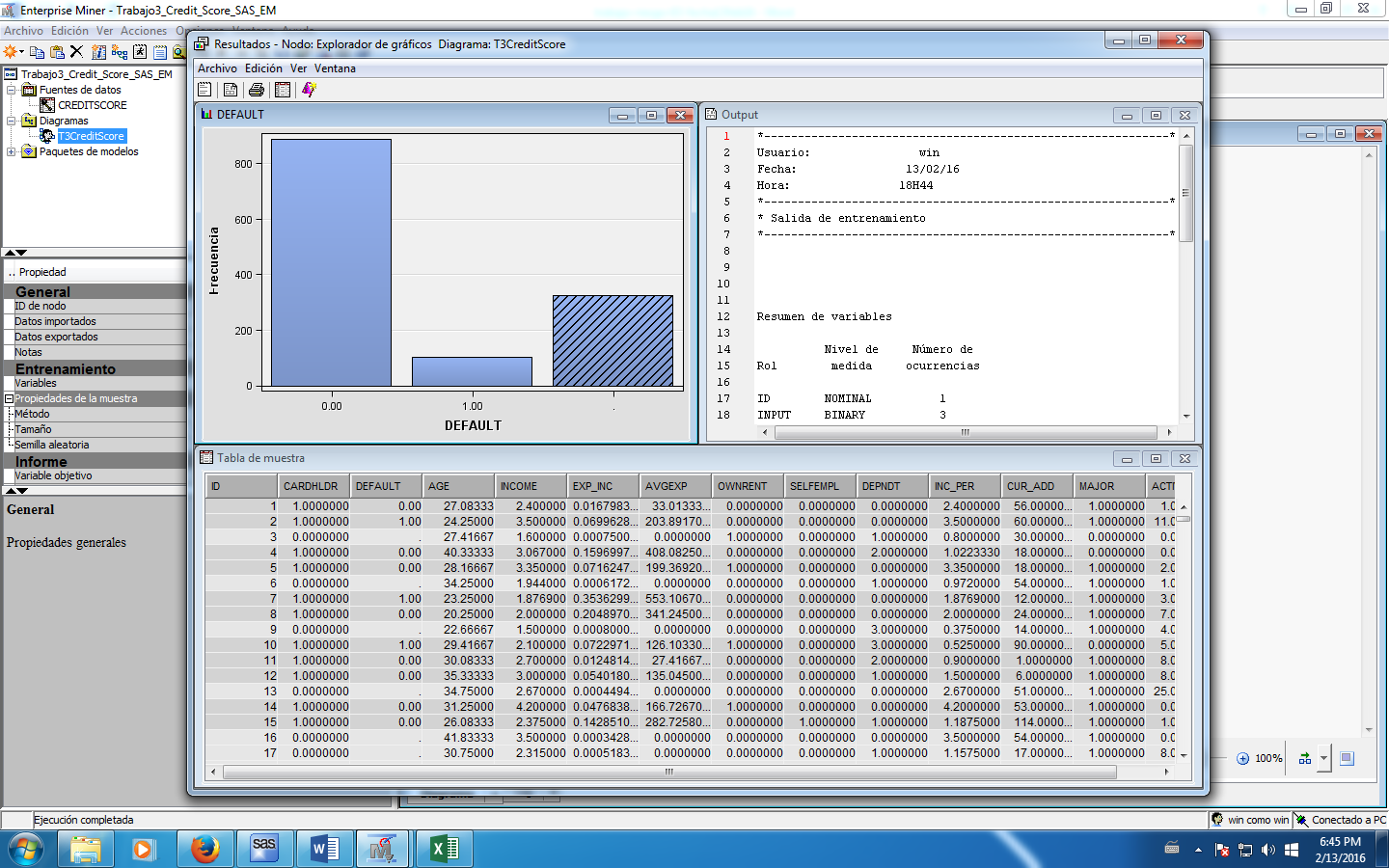
Default = 1 no ha pagado (malos) / 0 ha pagado (buenos)

En la imagen abajo se puede ver como se han quedado las variables mapeadas en binarias, intervalo y nominales.



**¿Qué porcentaje de buenos y malos tenemos en cada categoría?**

En la imagen abajo se puede ver la cantidad de default = 0 (Buenos) y default = 1 (malos) y Default = Vacío. Estos datos son para todos los datos.



Los datos **buenos** son los que tienen default = 0 y hay 890 observaciones.

Los datos **malos** son los que tienen default = 1 y hay 104 observaciones.

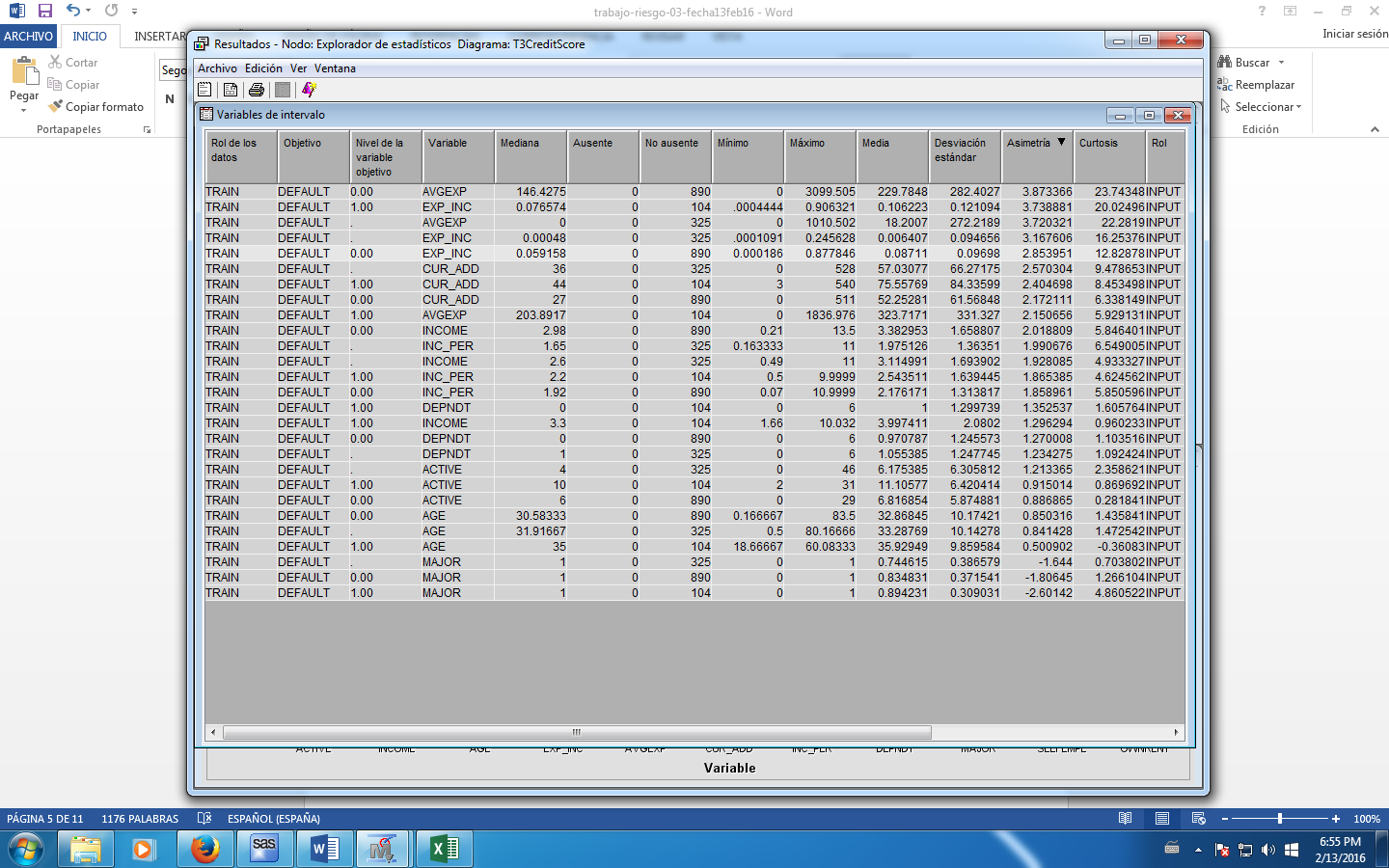
Los con default vacío son 325 observaciones.

**¿Hay datos anómalos?**

La variable edad (age) se puede ver un poco rara, no tiene valores redondos, se recomienda transformar.

Si se puede ver en algunas variables una gran diferencia en los valores de mínimo, máximo y media.

En la imagen abajo se puede ver las estadísticas para tener una idea de los datos.

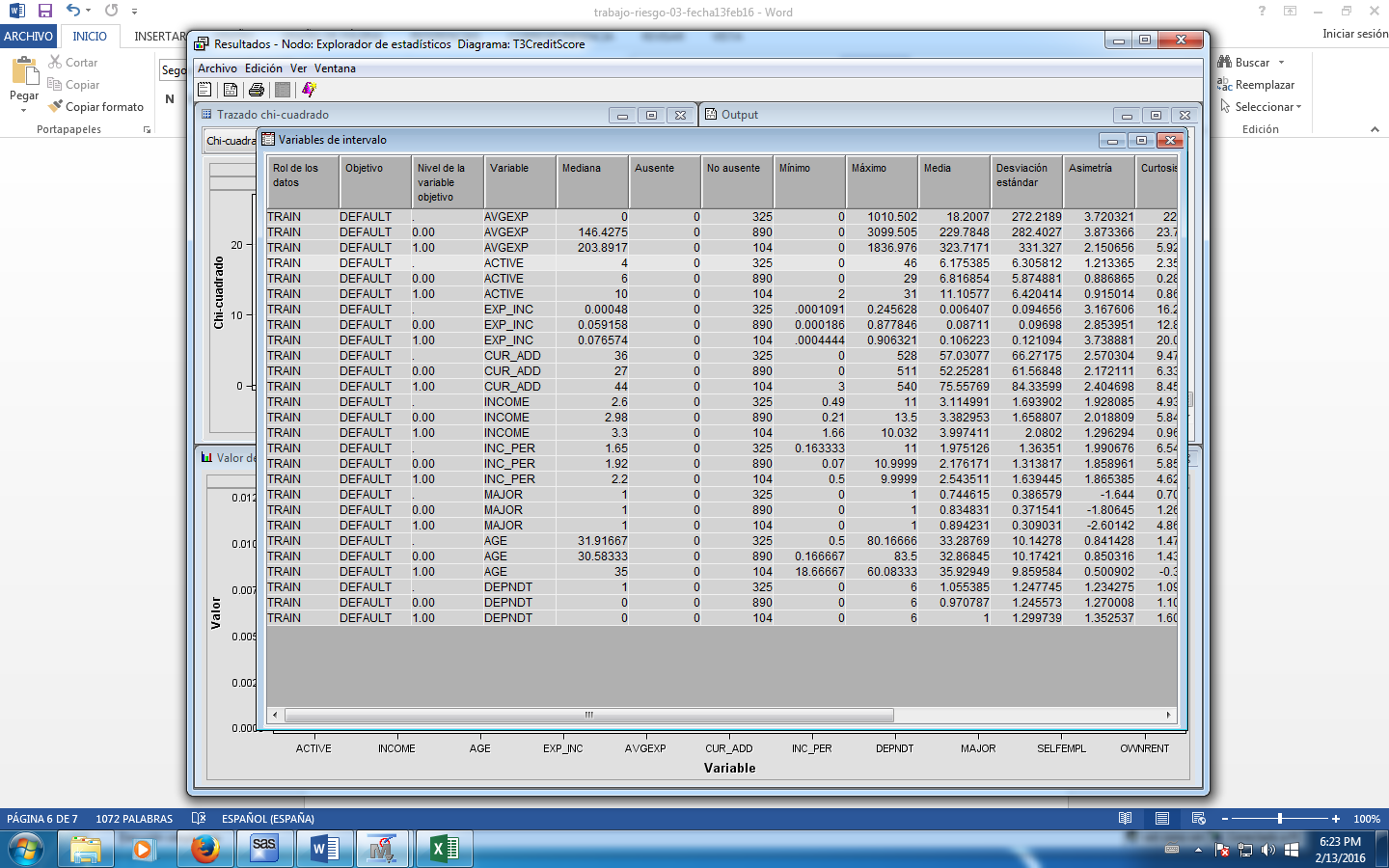


**¿Hay datos perdidos?**

**No hay datos nulos o perdidos** en el dataset. Solo hay datos nulos en las variables Cardhldr y default pero está correcto.

Cuando el cliente es nuevo no hay datos en Cardhldr y default.

Cuando el cliente ha sido rechazado y no ha recibido la tarjeta de crédito no hay datos en la columna default.



**¿Hay que imputar?**

*“En estadística, la imputación es la sustitución de valores no informados en una observación por otros.*

*A veces es un paso necesario para poder tratar los datos con determinadas técnicas estadísticas de análisis. Idealmente, este análisis debería tener en cuenta el hecho de que algunos de los datos no son observados sino que han sido imputados.” (Wikipedia).*

Yo he decidido no imputar datos.

1. **¿Hay que tramificar variables?**

La técnica de tramificar variables es transformar los valores en tramos.

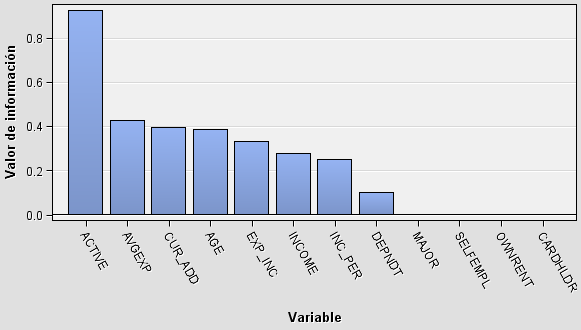
Ejemplo:

|  |  |
| --- | --- |
| Original | Tramificado |
| 1,2 | 1 |
| 3,8 | 4 |

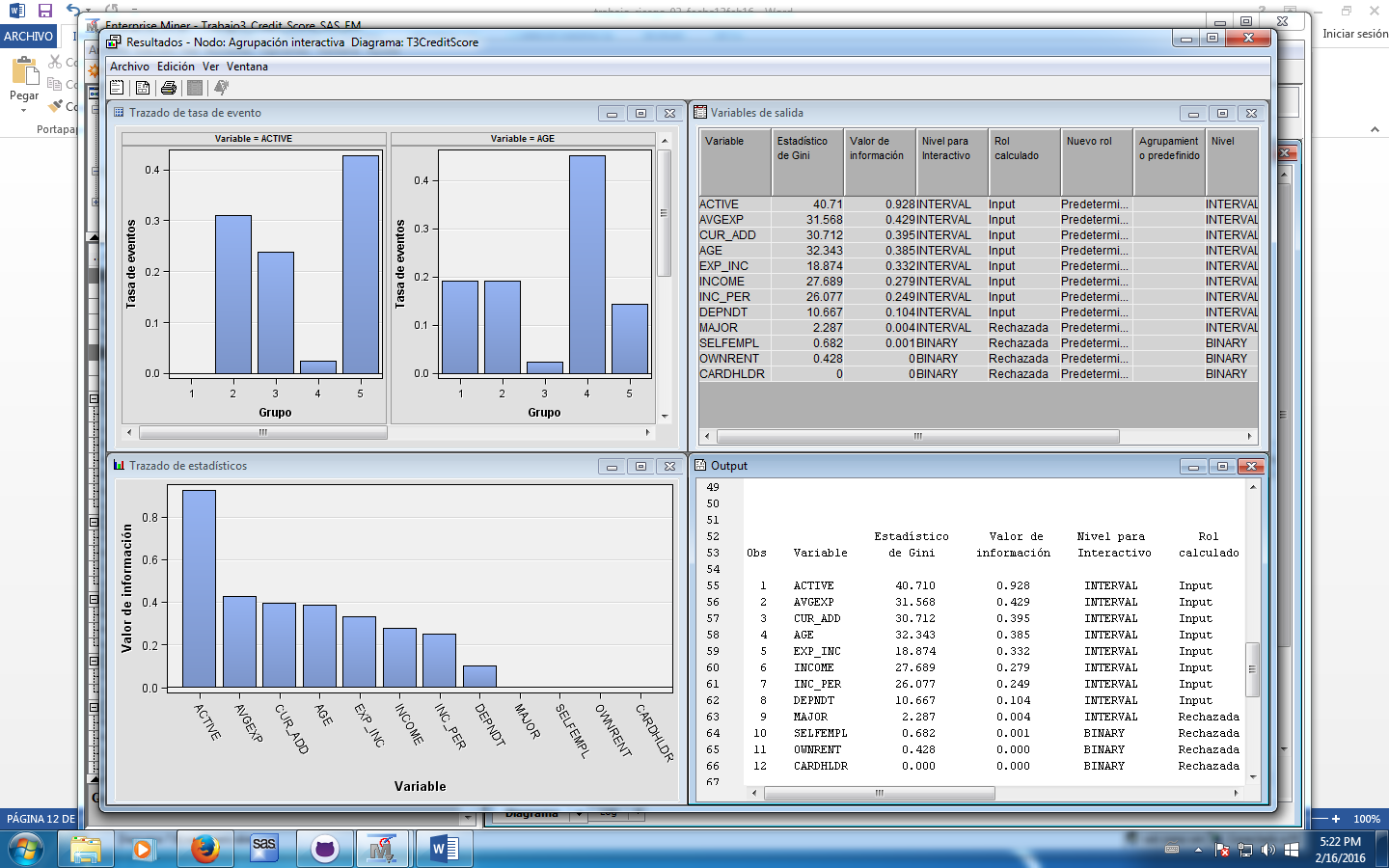
Yo no he tramificado ninguna variable del dataset, pero quizás se puede recomendar tramificar la variable edad (age) que tiene valores no enteros.

1. **Selección de variables ¿qué variables son las más importantes?**

Utilizando el nodo “Agrupación interactiva” se tiene las variables más importantes por valor de información como se puede ver en la imagen abajo.

****

En la imagen abajo se puede ver también las tablas con los estadísticos de Gini y valor de información.



**¿qué variables deben incluirse en el scorecard?**

Yo he puesto todas las variables en el scorecard, pero se podría quitar las que MAJOR, SELFEMPL, OWNRENT y CARDHLDR.

Son ellas:

* ID = Identificador
* Cardhldr = 1 ha recebido la tarjeta / 0 clientes rechazados
* Default = 1 no ha pagado (malos) / 0 ha pagado (buenos)
* Age = Age n years plus twelfths of a year
* Income = Yearly income (divided by 10,000)
* Exp\_Inc = Ratio of monthly credit card expenditure to yearly income
* Avgexp = Average monthly credit card expenditure
* Ownrent = 1 if owns their home, 0 if rent
* Selfempl = 1 if self employed, 0 if not.
* Depndt = 1 + number of dependents
* Inc\_per = Income divided by number of dependents
* Cur\_add = months living at current address
* Major = number of major credit cards held
* Active = number of active credit accounts

Abajo una tabla con las variables y el tipo de cada variable definido por mí.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Tipo** | **Rol** |
| ID | Nominal | ID |
| Cardhldr | Binaria | Input |
| Default | Binaria | Objetivo |
| Age | Intervalo | Input |
| Income | Intervalo | Input |
| Exp\_Inc | Intervalo | Input |
| Avgexp | Intervalo | Input |
| Ownrent | Binaria | Input |
| Selfempl | Binaria | Input |
| Depndt | Intervalo | Input |
| Inc\_per | Intervalo | Input |
| Cur\_add | Intervalo | Input |
| Major | Intervalo | Input |
| Active | Invervalo | Input |

Disponibilidad, e importancia, estadísticos de selección (IV, correlación)

Para ver la correlación en SAS:

**data** cs\_accepts;

set riesgo.Cs\_accepts;

**run**;

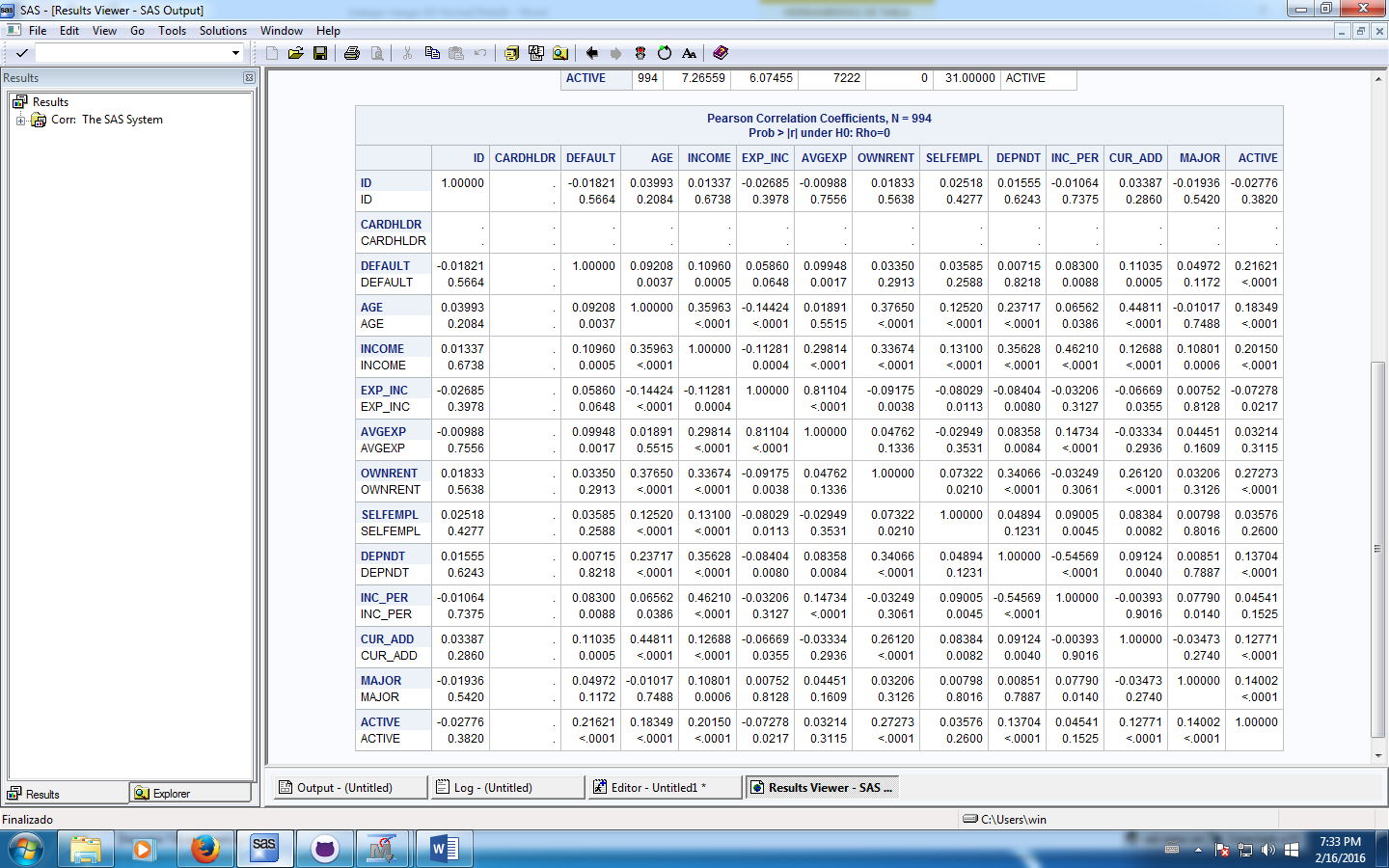
**proc** **corr** data=cs\_accepts outp=cs\_accepts\_corr;

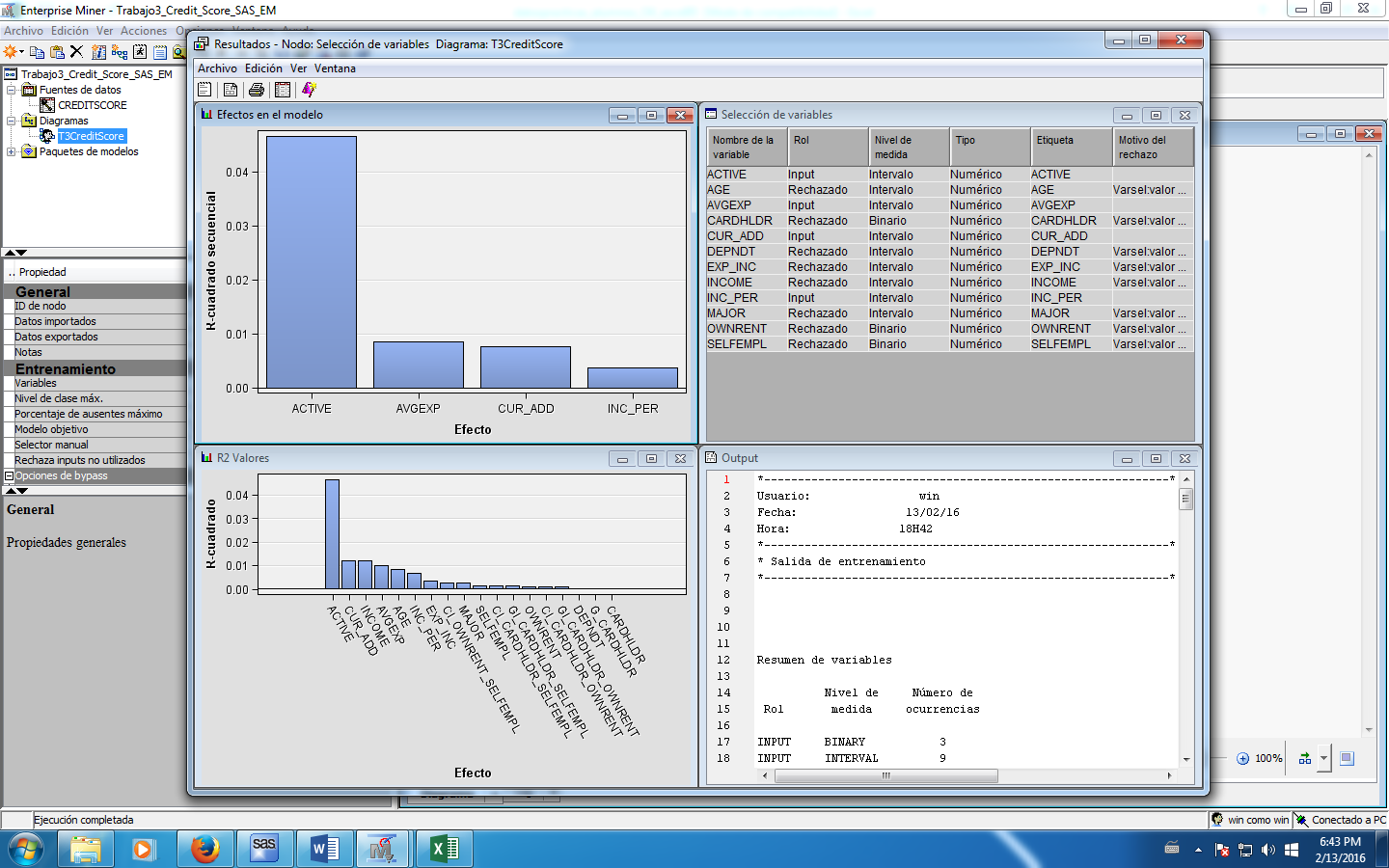
**run**;

The CORR Procedure

|  |  |
| --- | --- |
| **14 Variables:** | ID CARDHLDR DEFAULT AGE INCOME EXP\_INC AVGEXP OWNRENT SELFEMPL DEPNDT INC\_PER CUR\_ADD MAJOR ACTIVE |

| **Simple Statistics** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** | **Label** |
| **ID** | 994 | 643.06237 | 366.55474 | 639204 | 1.00000 | 1285 | ID |
| **CARDHLDR** | 994 | 1.00000 | 0 | 994.00000 | 1.00000 | 1.00000 | CARDHLDR |
| **DEFAULT** | 994 | 0.10463 | 0.30623 | 104.00000 | 0 | 1.00000 | DEFAULT |
| **AGE** | 994 | 33.18872 | 10.18014 | 32990 | 0.16667 | 83.50000 | AGE |
| **INCOME** | 994 | 3.44724 | 1.71689 | 3427 | 0.21000 | 13.50000 | INCOME |
| **EXP\_INC** | 994 | 0.08911 | 0.09988 | 88.57550 | 0.0001860 | 0.90632 | EXP\_INC |
| **AVGEXP** | 994 | 239.61272 | 289.15890 | 238175 | 0 | 3100 | AVGEXP |
| **OWNRENT** | 994 | 0.47988 | 0.49985 | 477.00000 | 0 | 1.00000 | OWNRENT |
| **SELFEMPL** | 994 | 0.06137 | 0.24013 | 61.00000 | 0 | 1.00000 | SELFEMPL |
| **DEPNDT** | 994 | 0.97384 | 1.25071 | 968.00000 | 0 | 6.00000 | DEPNDT |
| **INC\_PER** | 994 | 2.21460 | 1.35528 | 2201 | 0.07000 | 10.99990 | INC\_PER |
| **CUR\_ADD** | 994 | 54.69115 | 64.67114 | 54363 | 0 | 540.00000 | CUR\_ADD |
| **MAJOR** | 994 | 0.84105 | 0.36582 | 836.00000 | 0 | 1.00000 | MAJOR |
| **ACTIVE** | 994 | 7.26559 | 6.07455 | 7222 | 0 | 31.00000 | ACTIVE |

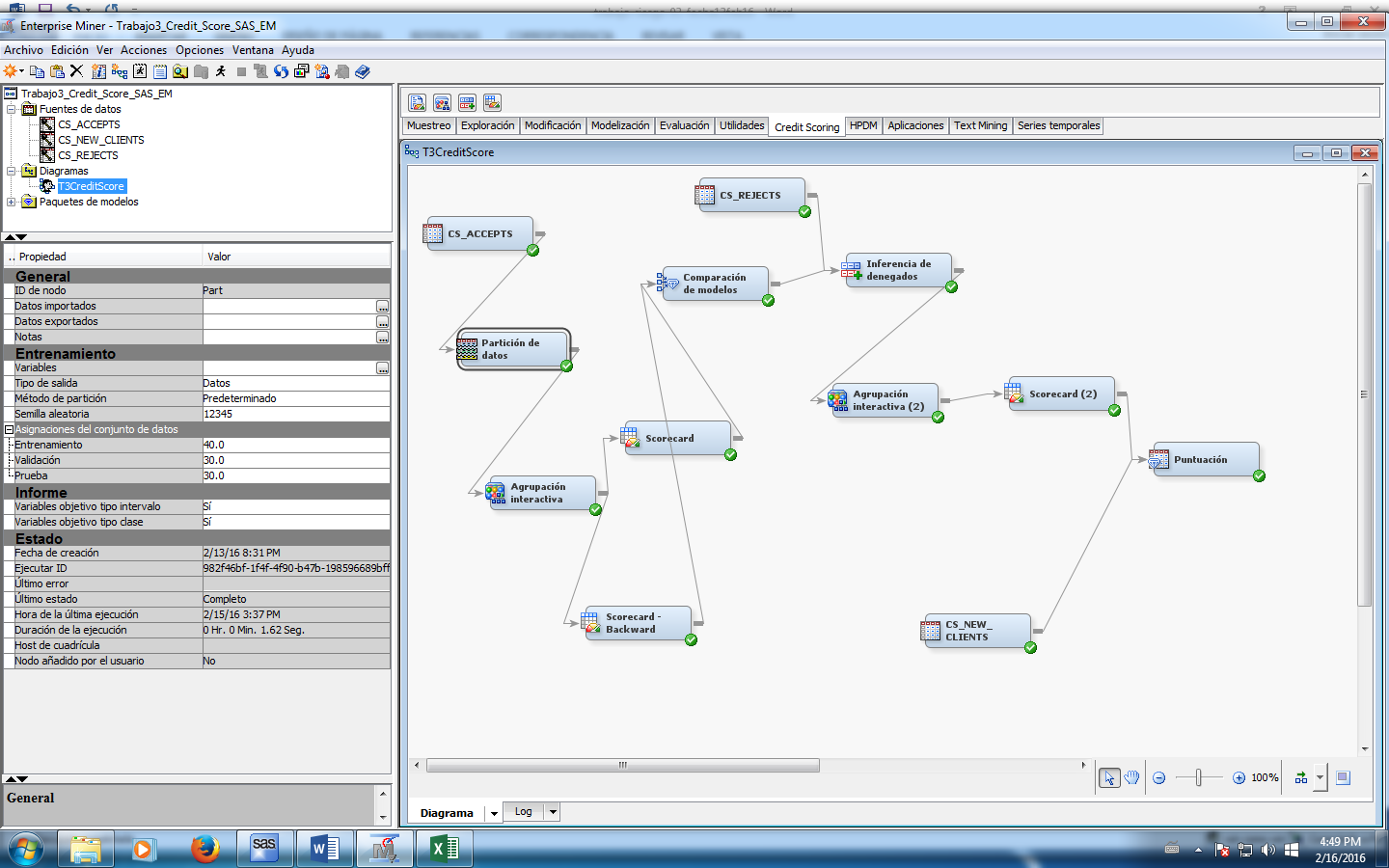


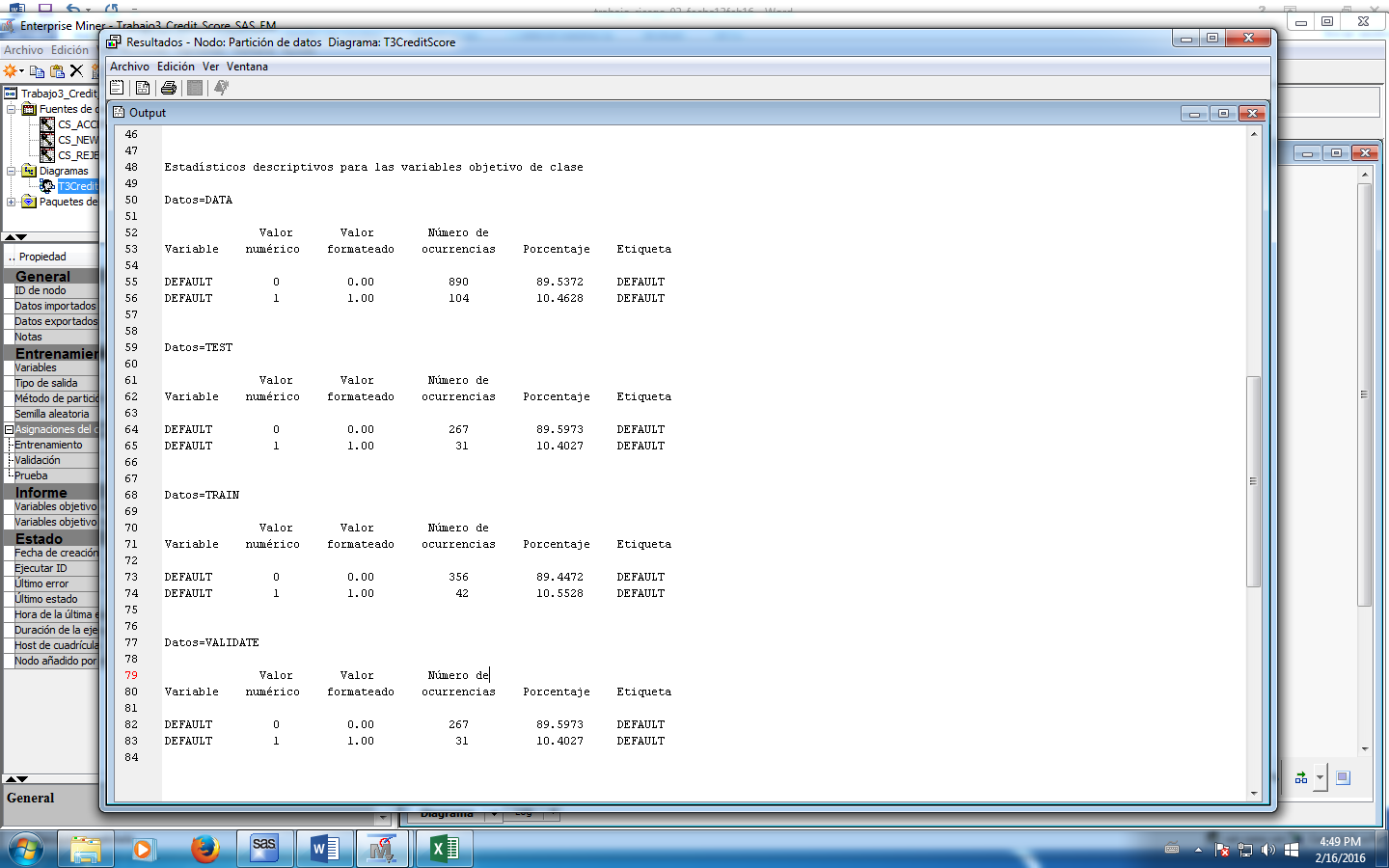


**5) Muestreo:**

Muestra de desarrollo/entrenamiento (70-80%) y de validación (30-20%)

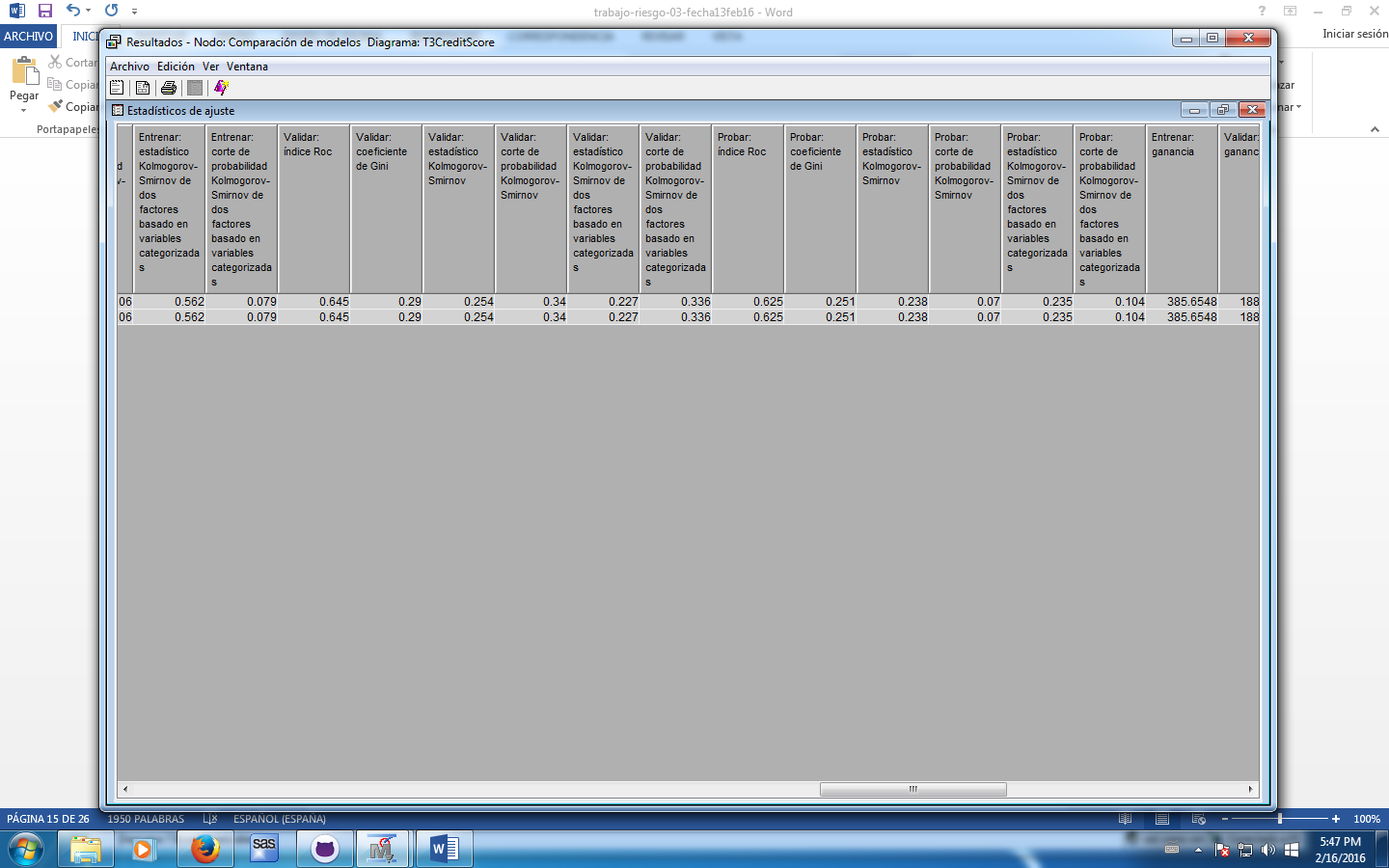
Sobre muestreo de buenos/malos?

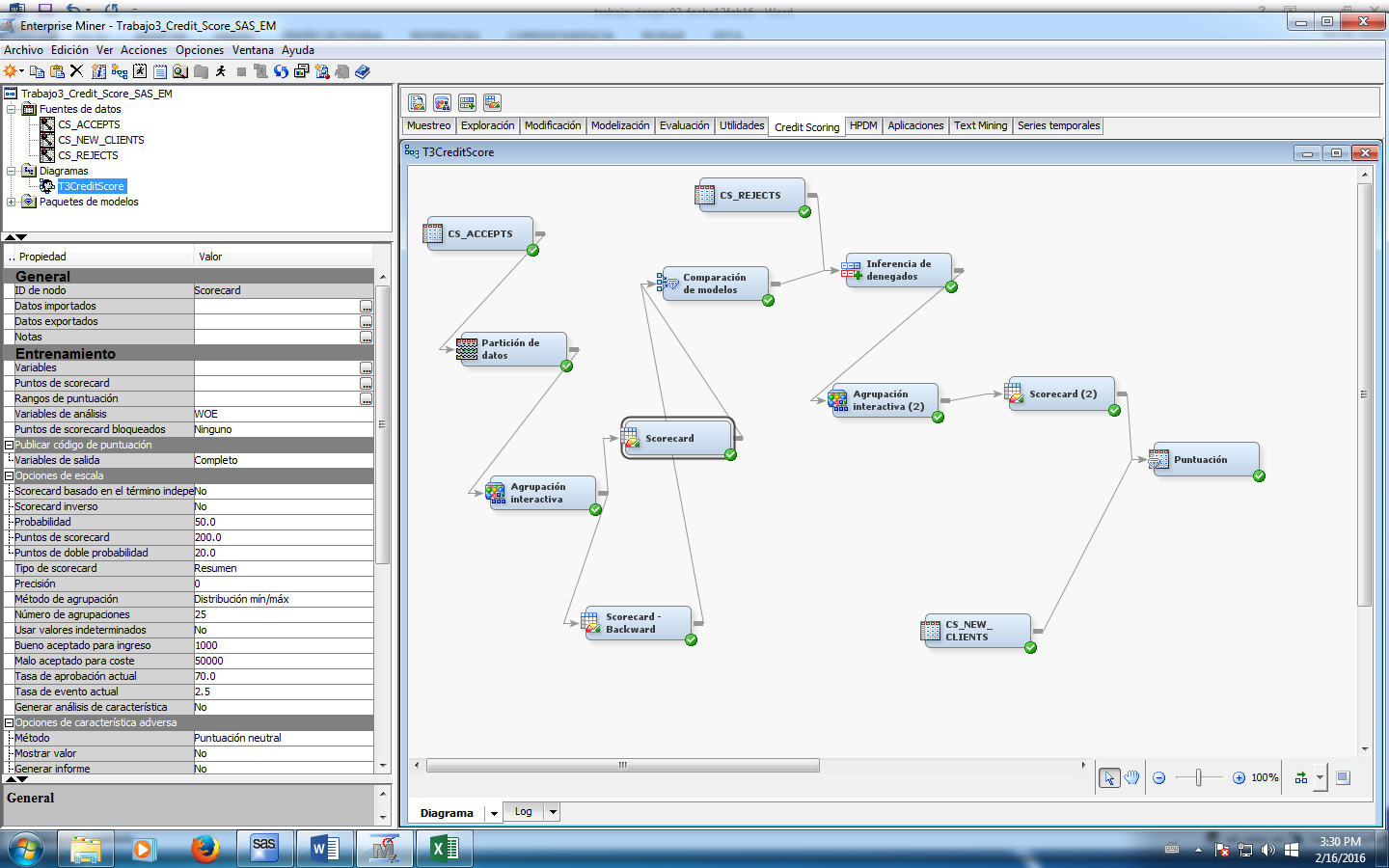


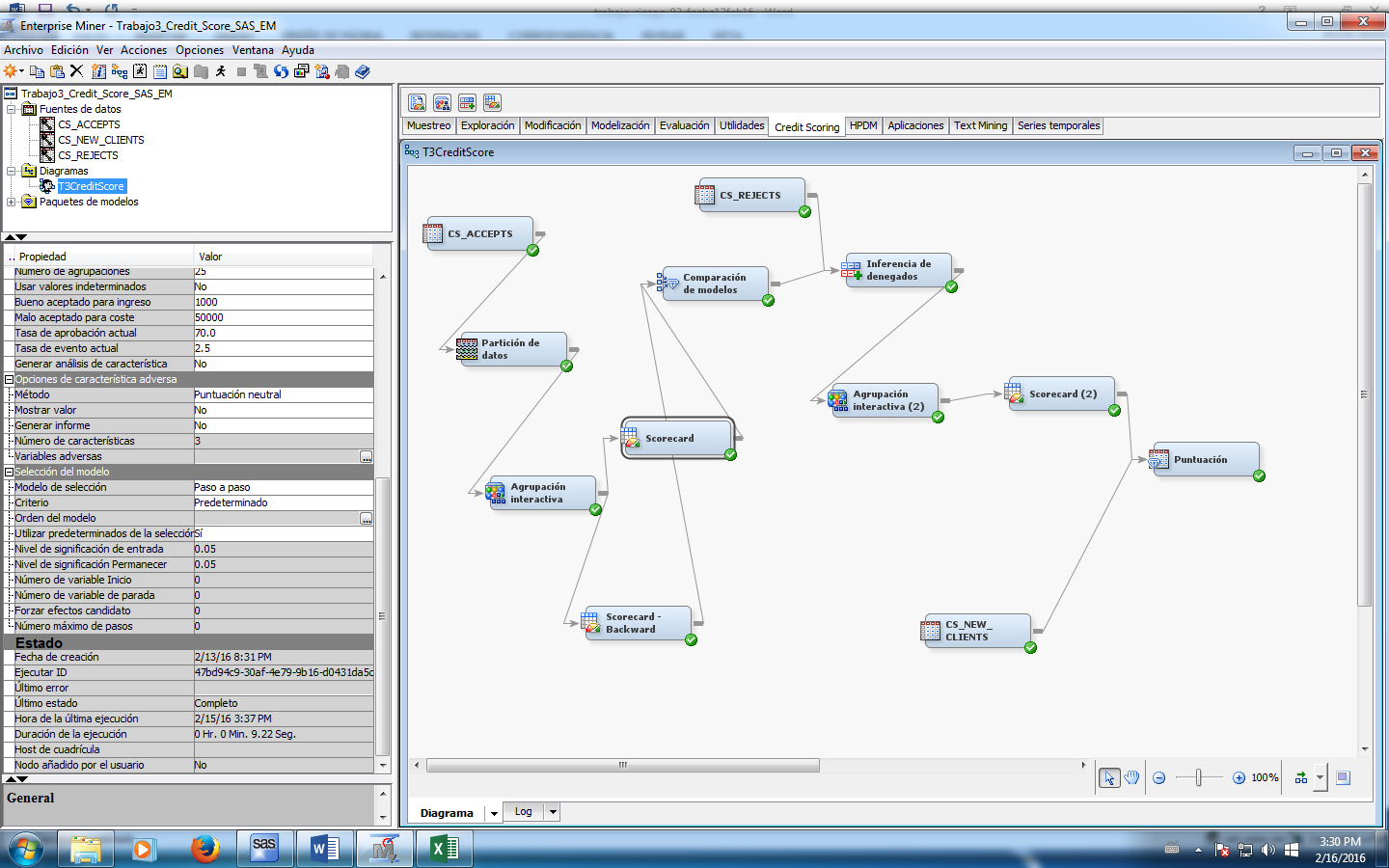


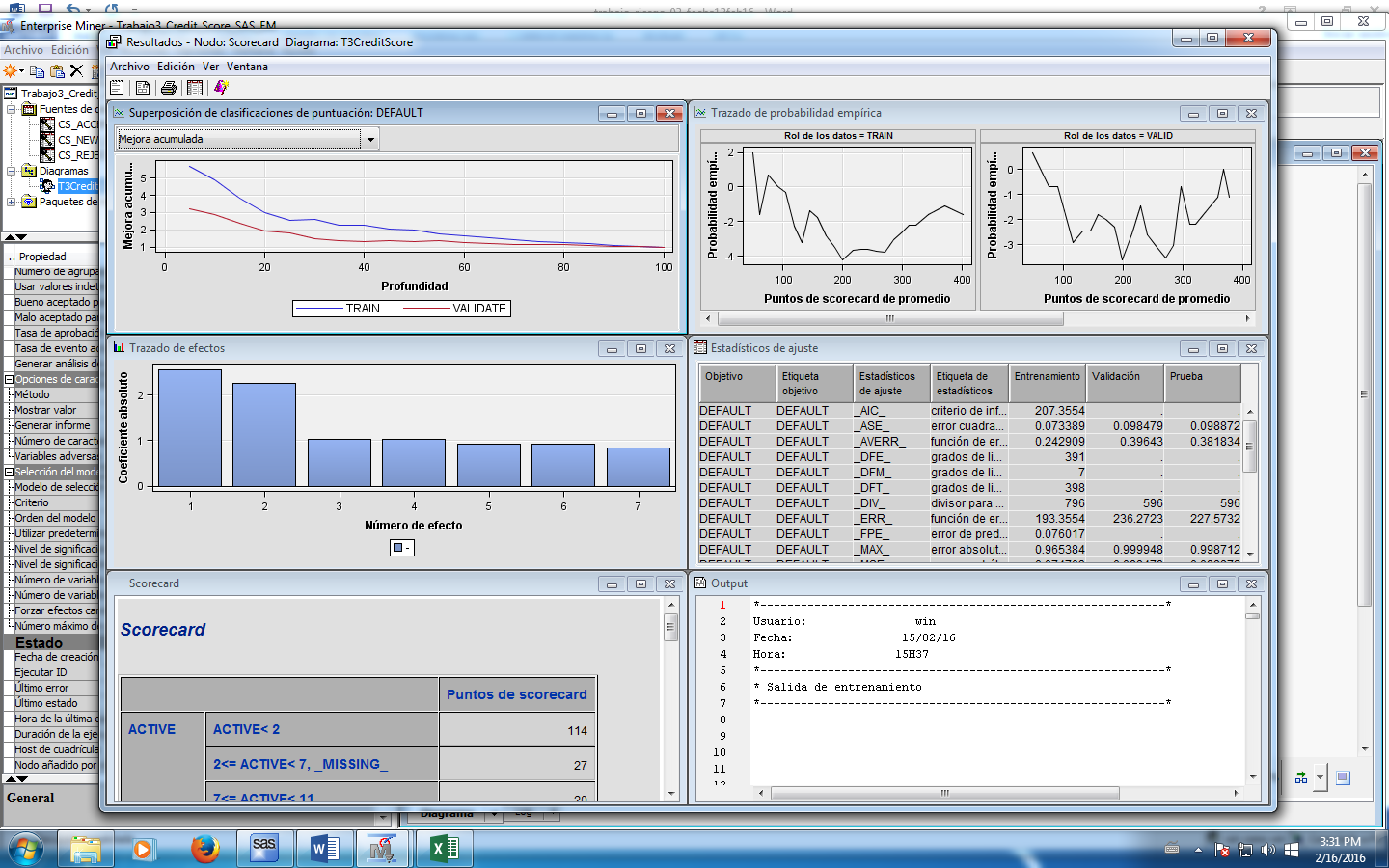
**6) Estimación del primer modelo, diagnosis (Kolmogorov-Smirnov, c-statistic, Gini) y obtención del scorecard preliminar**

En la imagen abajo se puede ver el diagnosis del primer modelo.

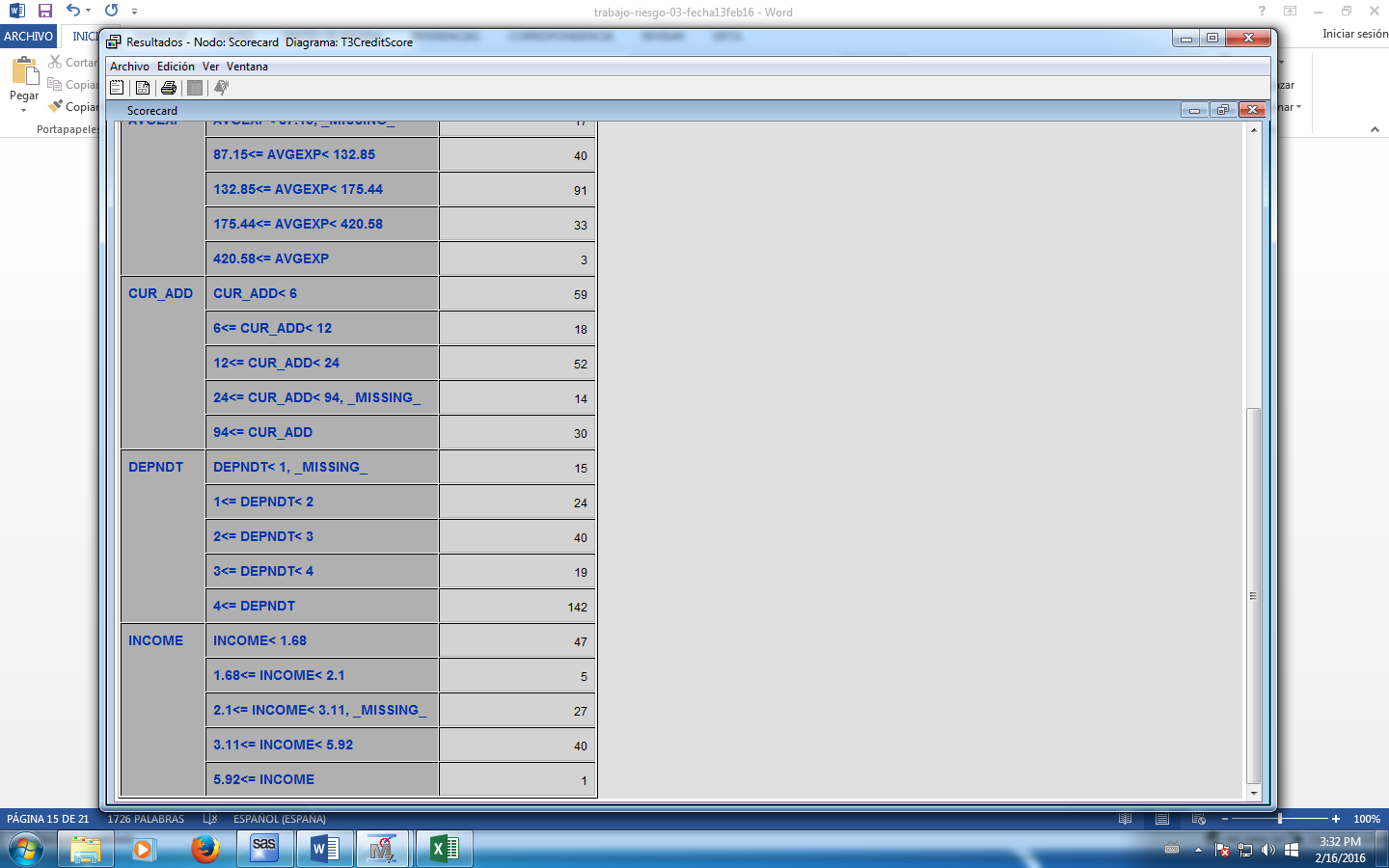






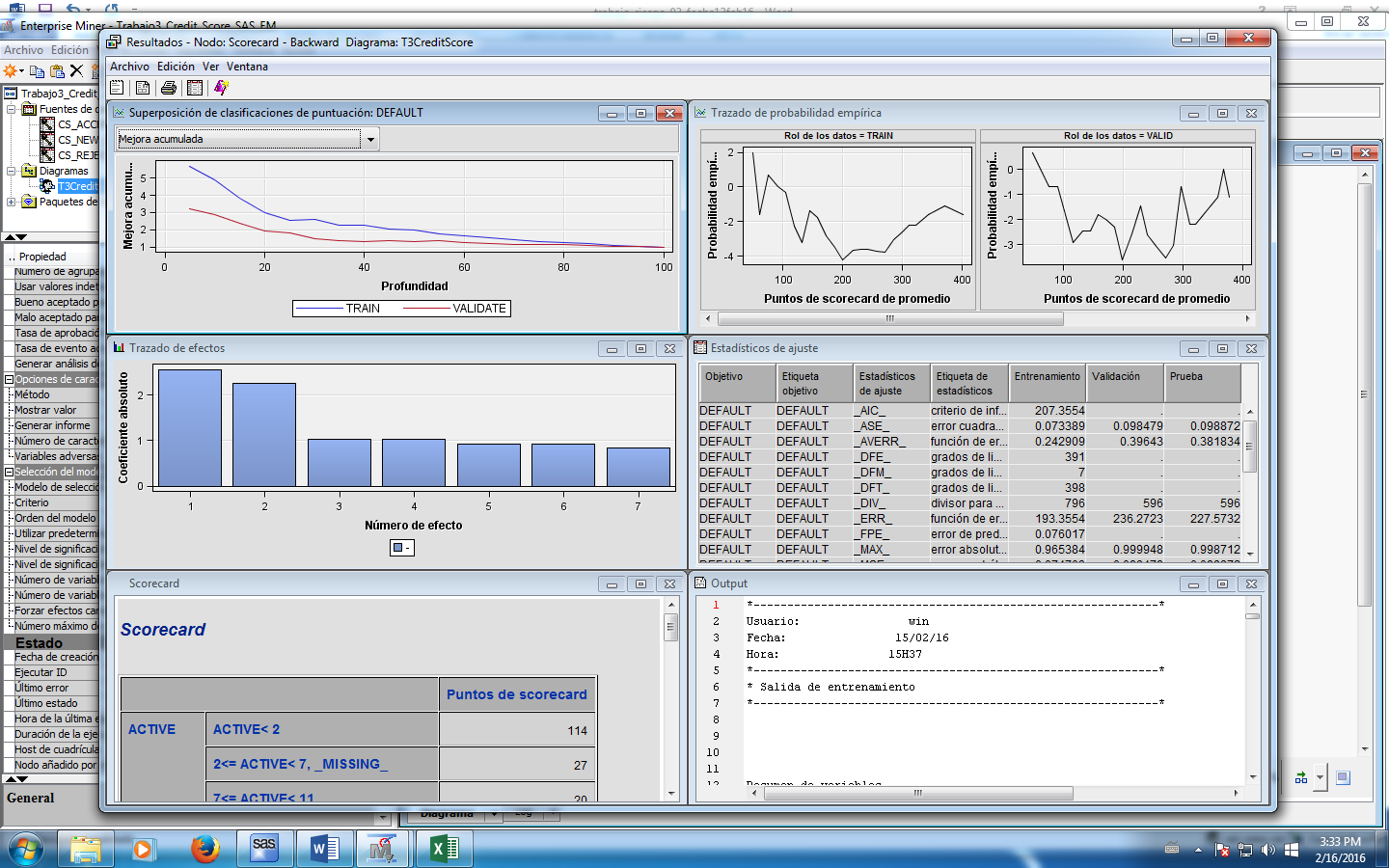




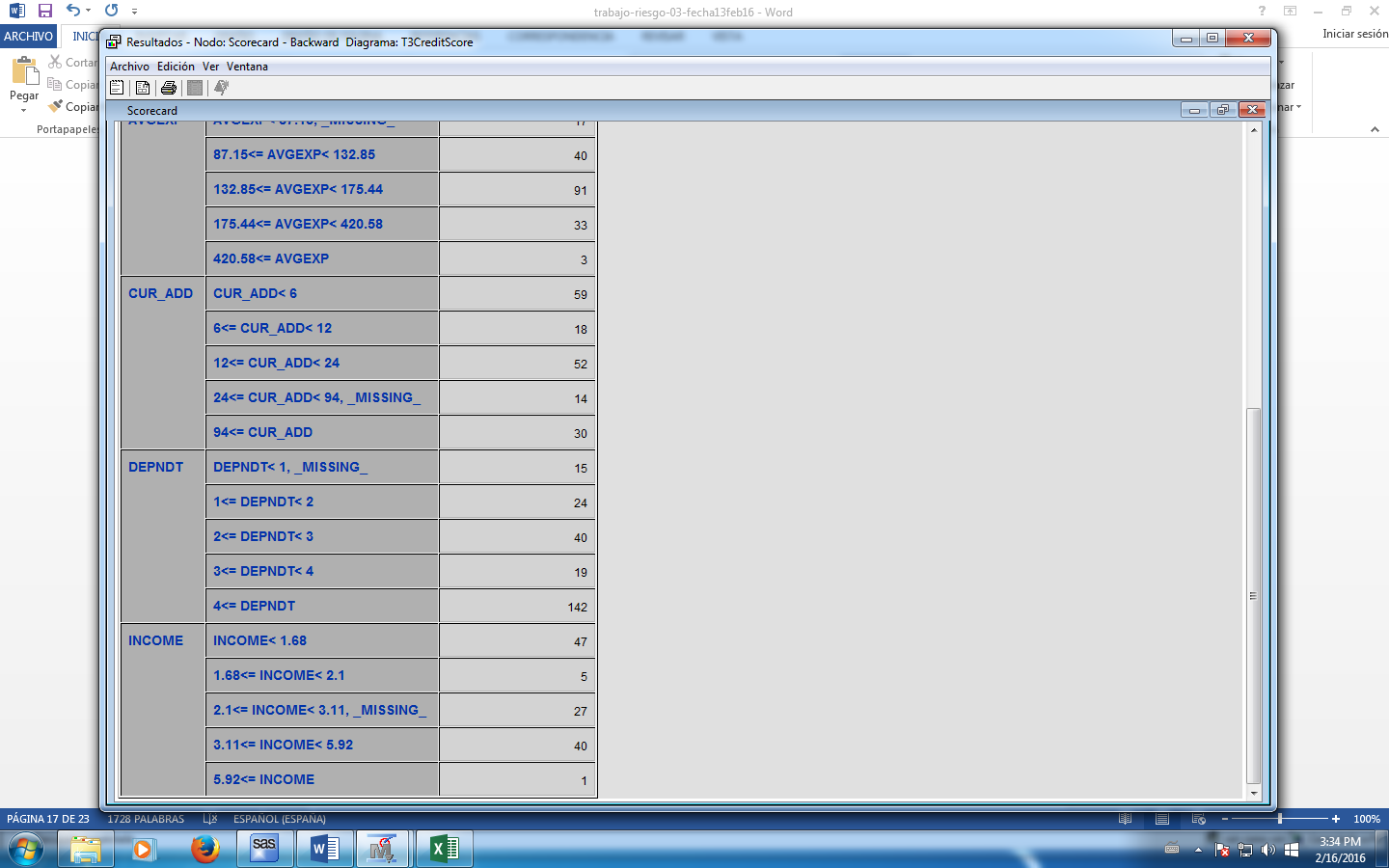


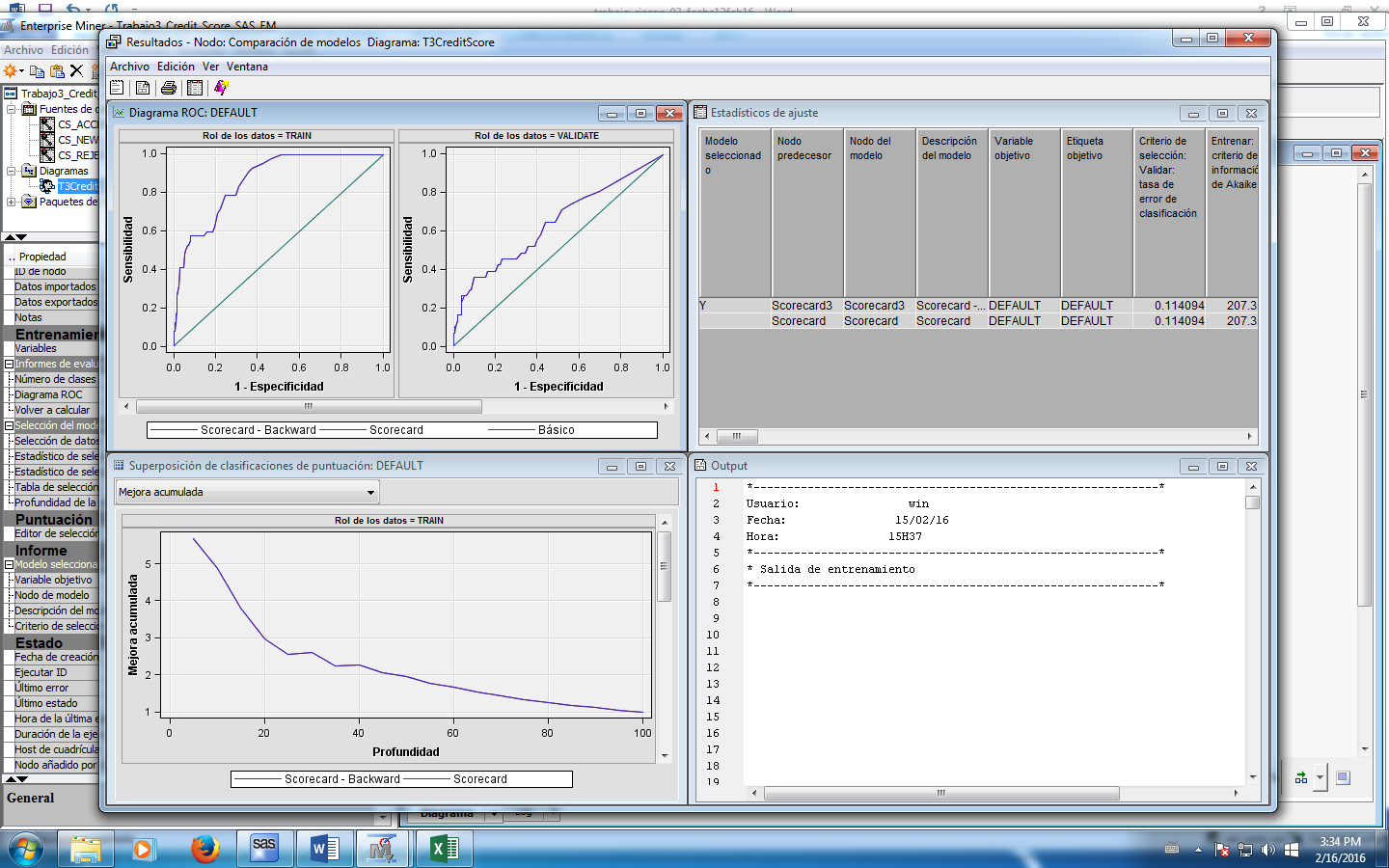
Scorecard Backward



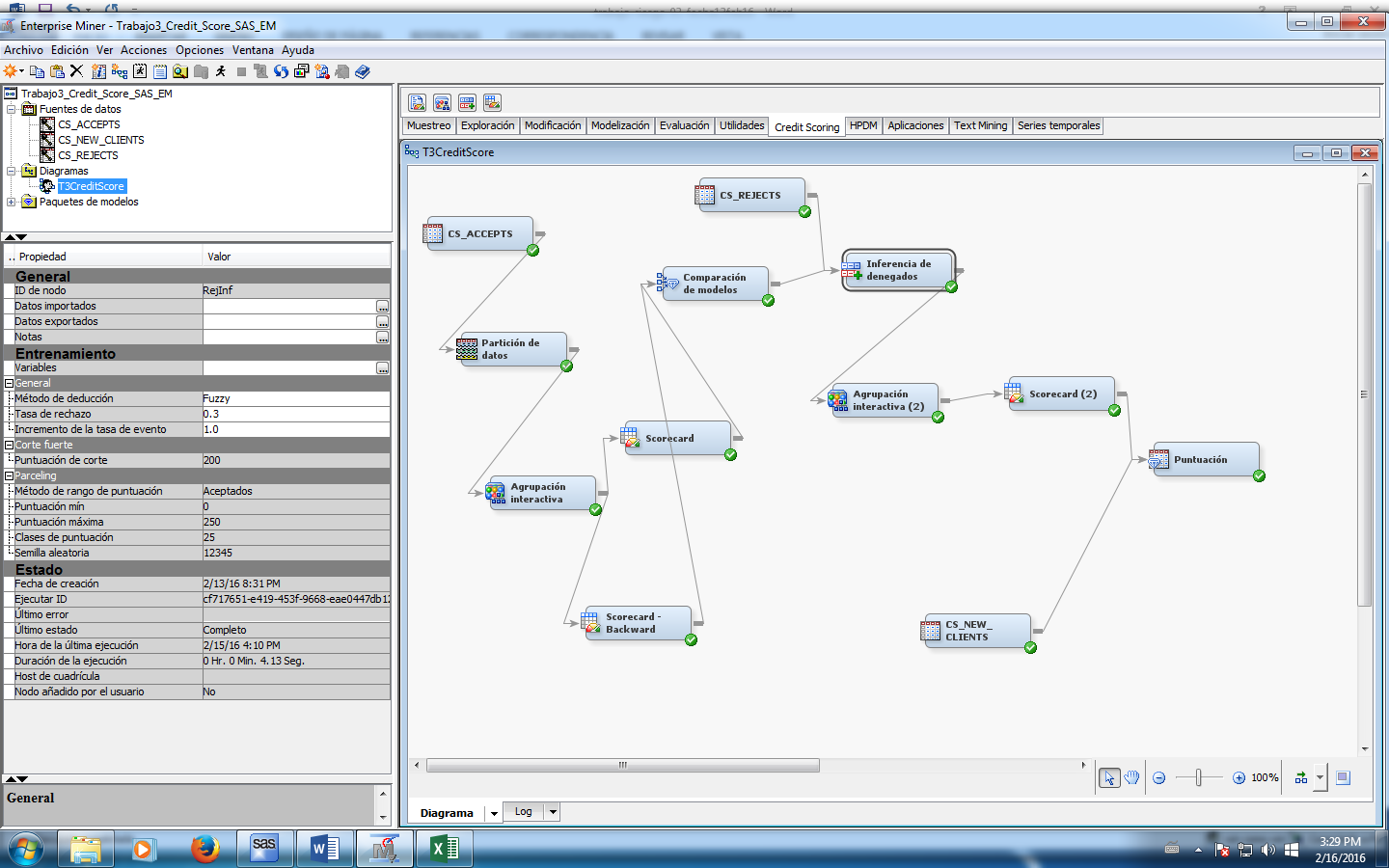


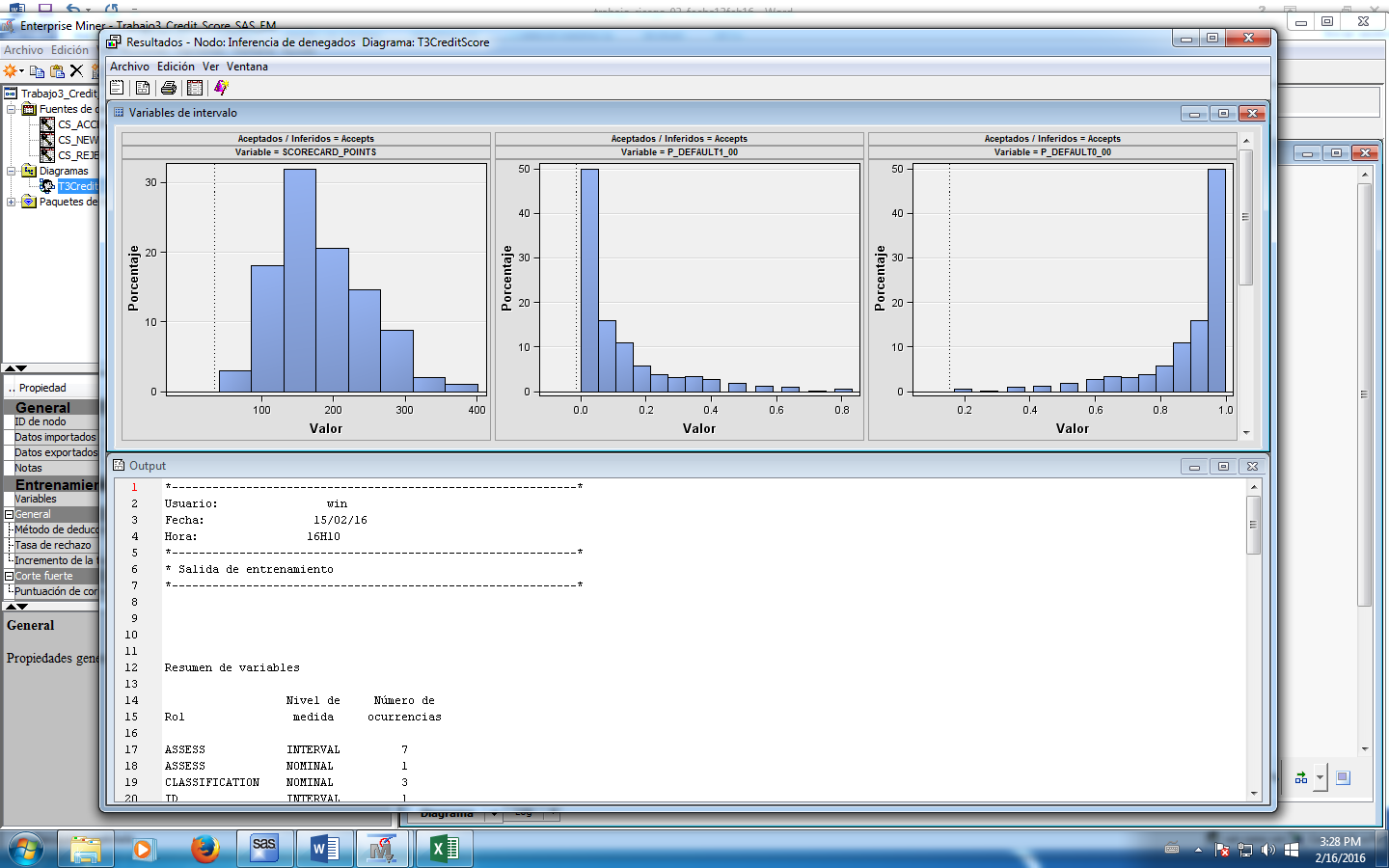




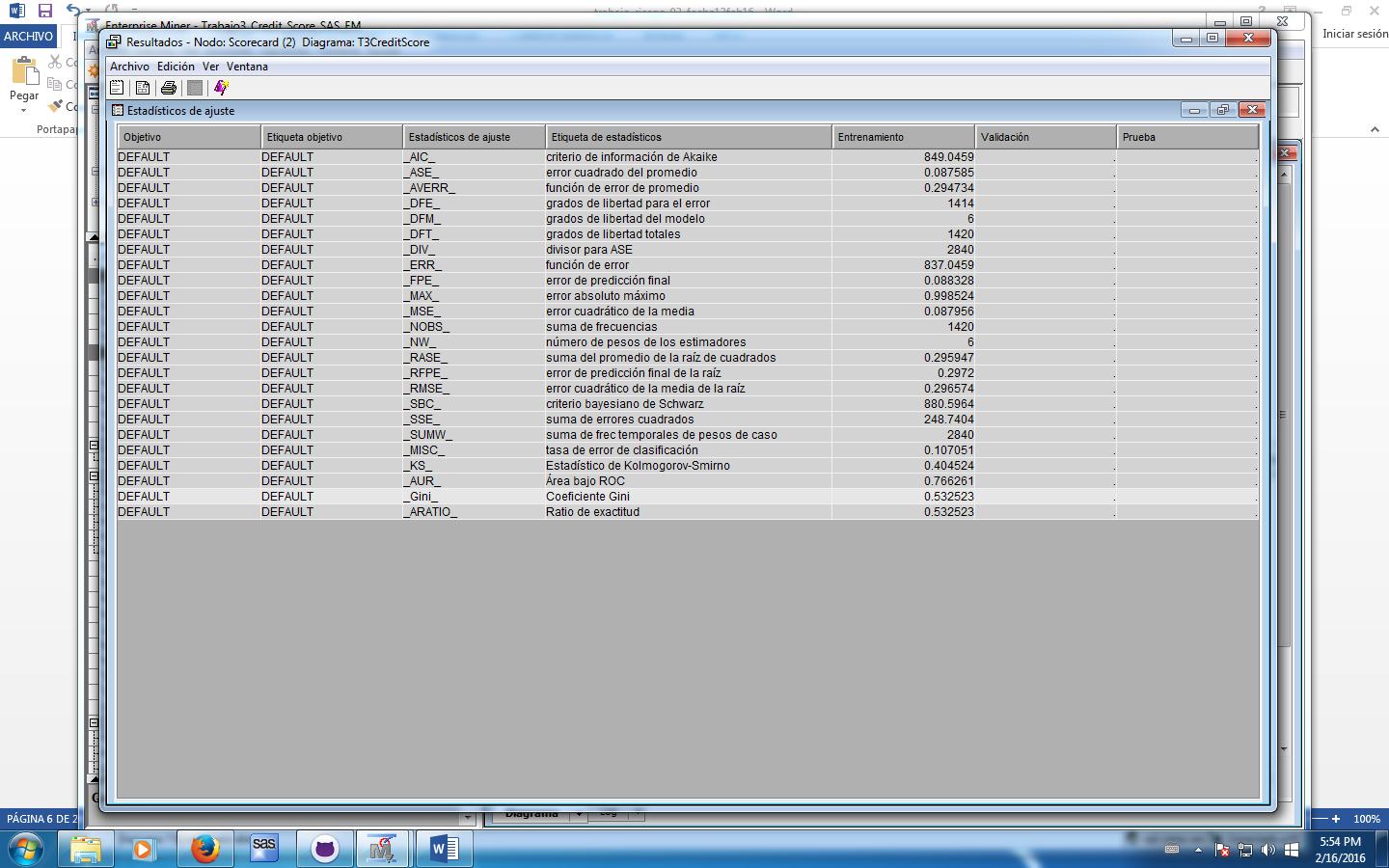


**7) Inferencia de denegados**



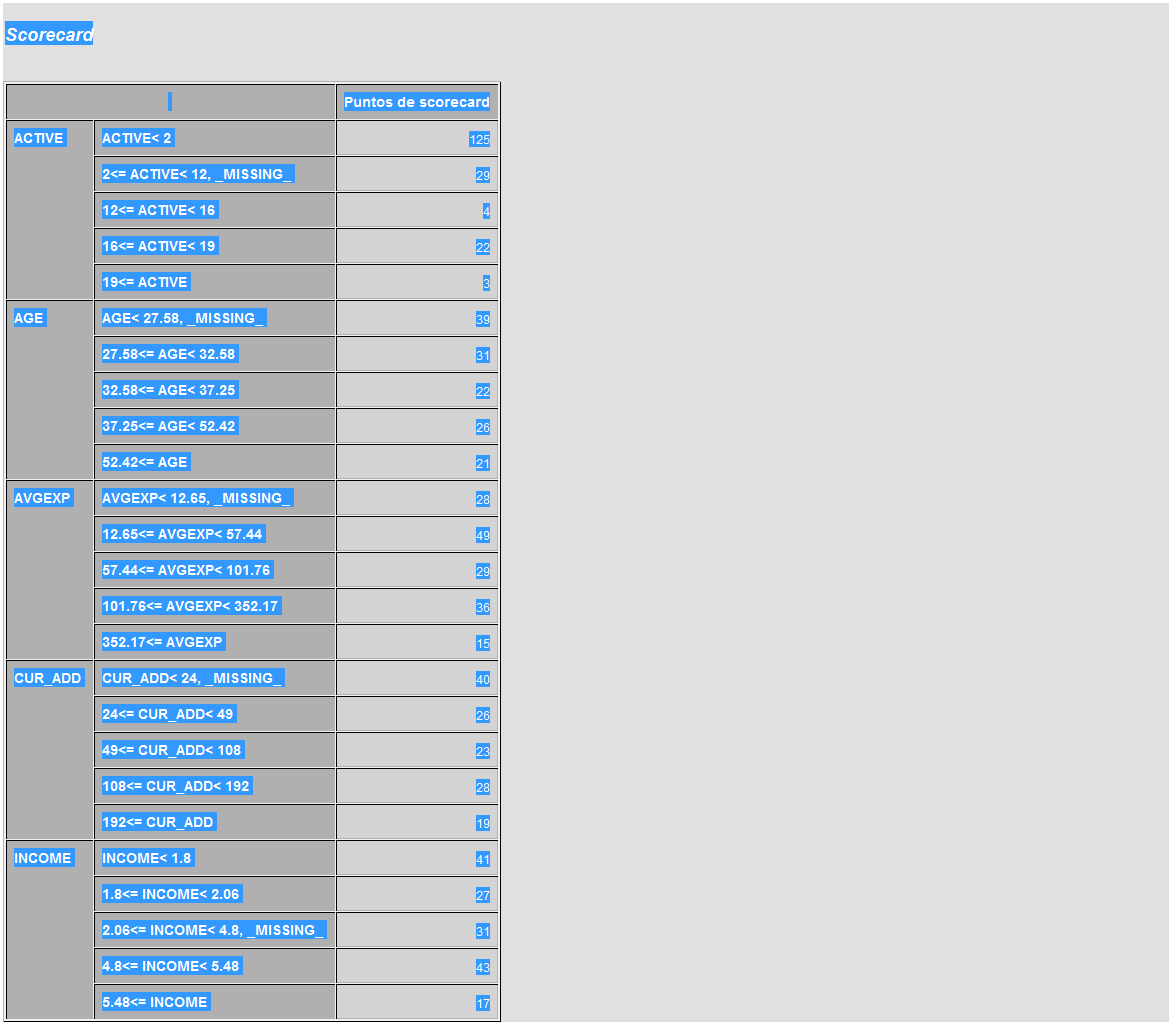


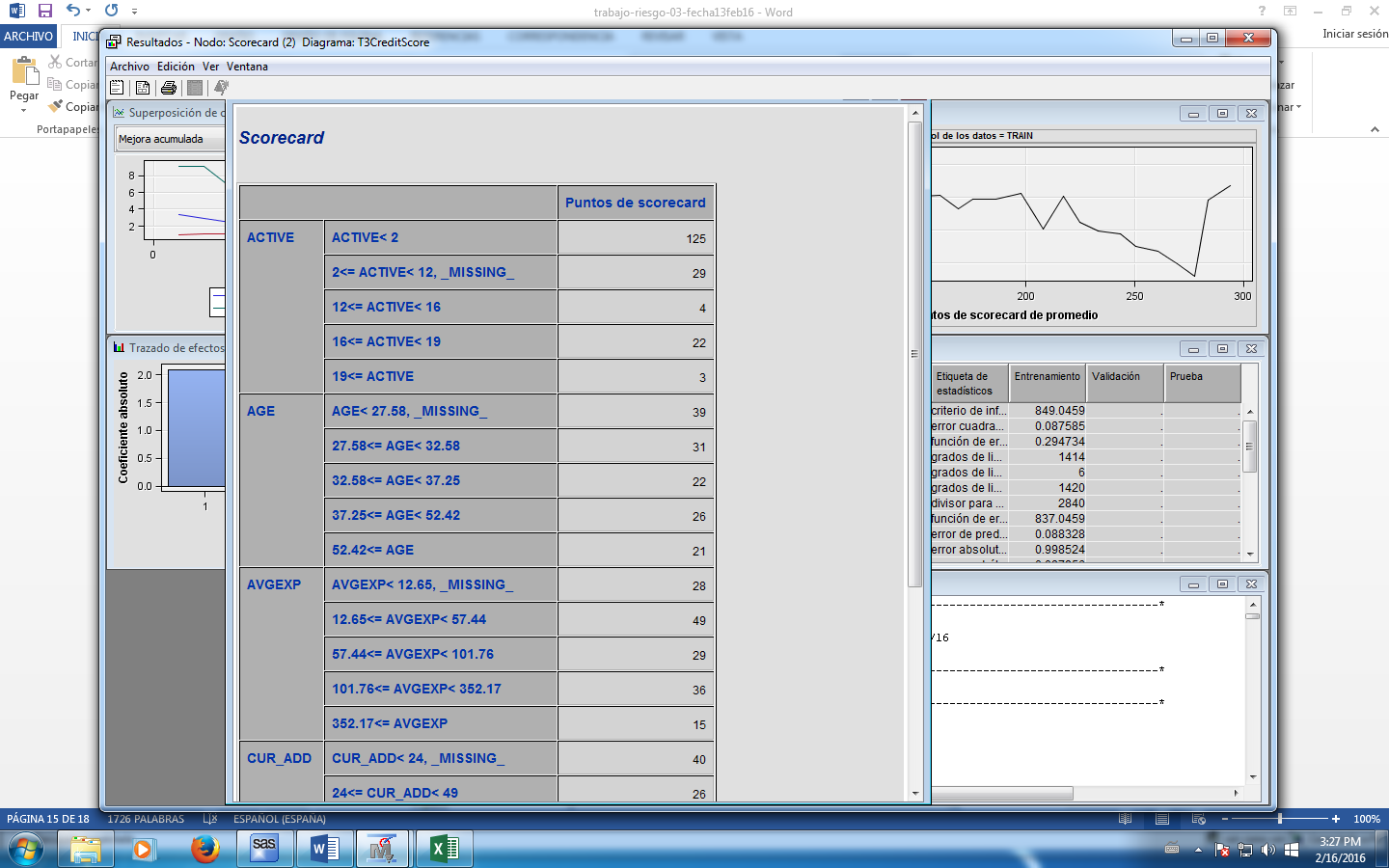
**8) Estimación del modelo definitivo y obtención del scorecard definitivo**





Scorecard definitivo

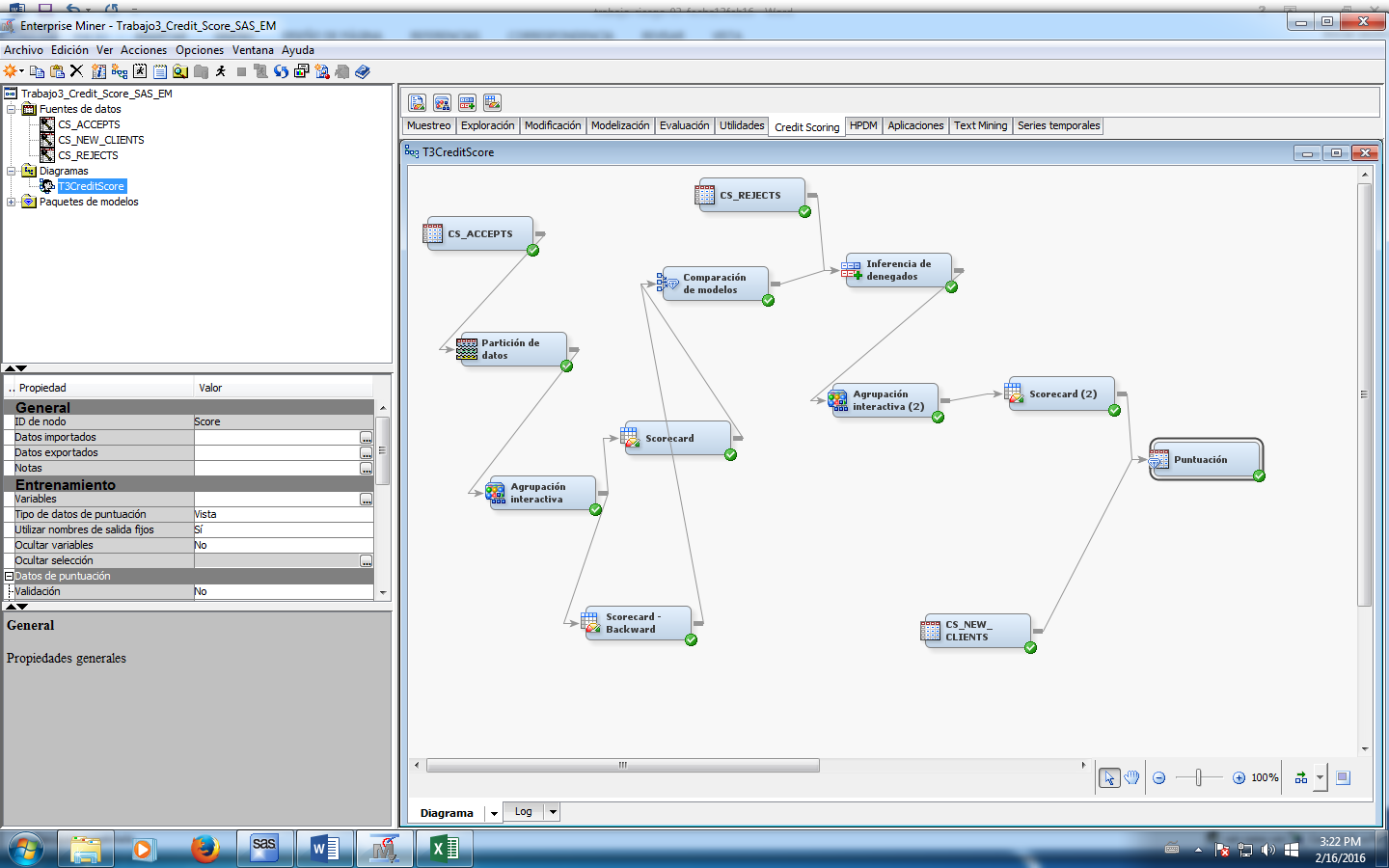




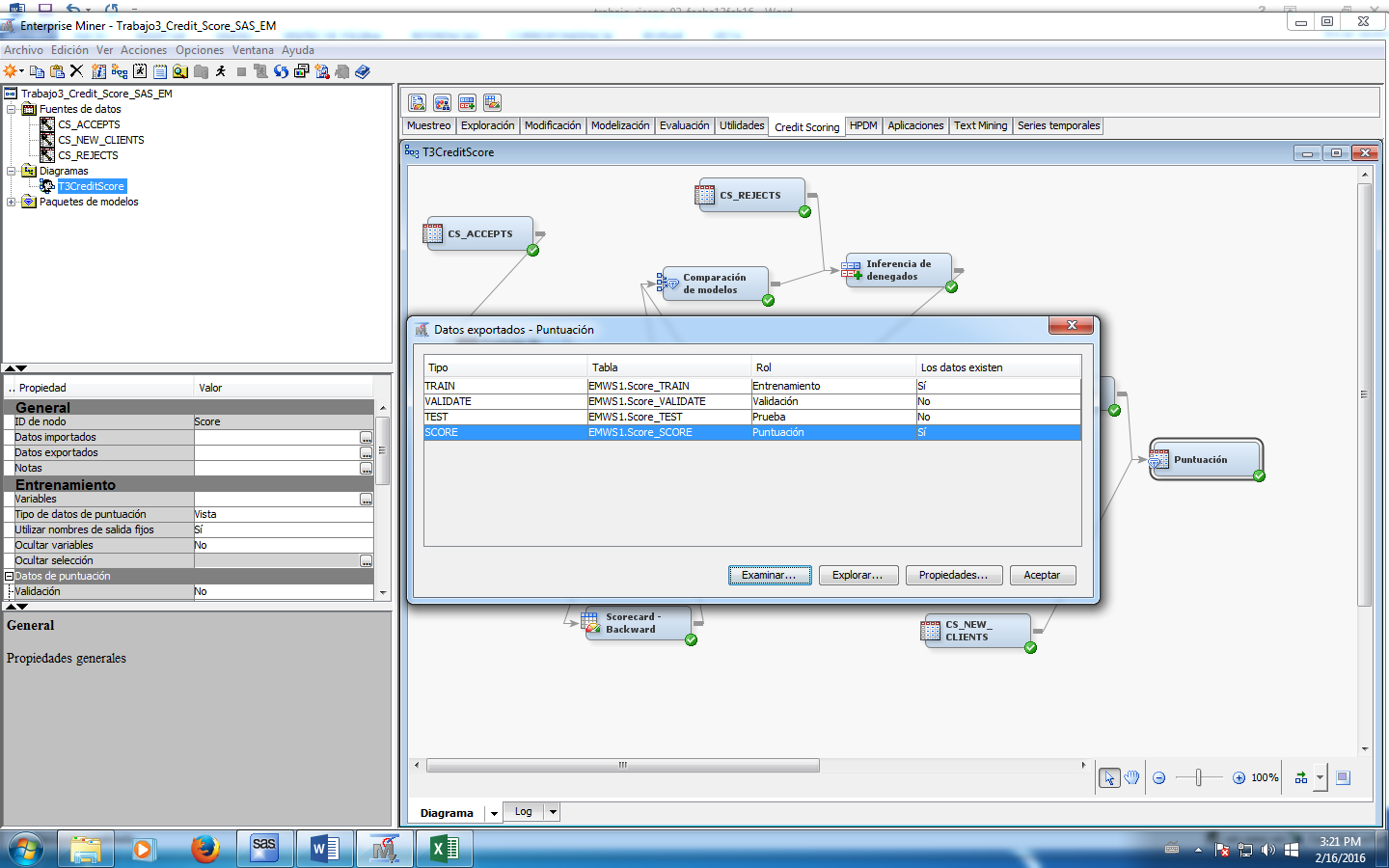


**9) Validación y Seguimiento**

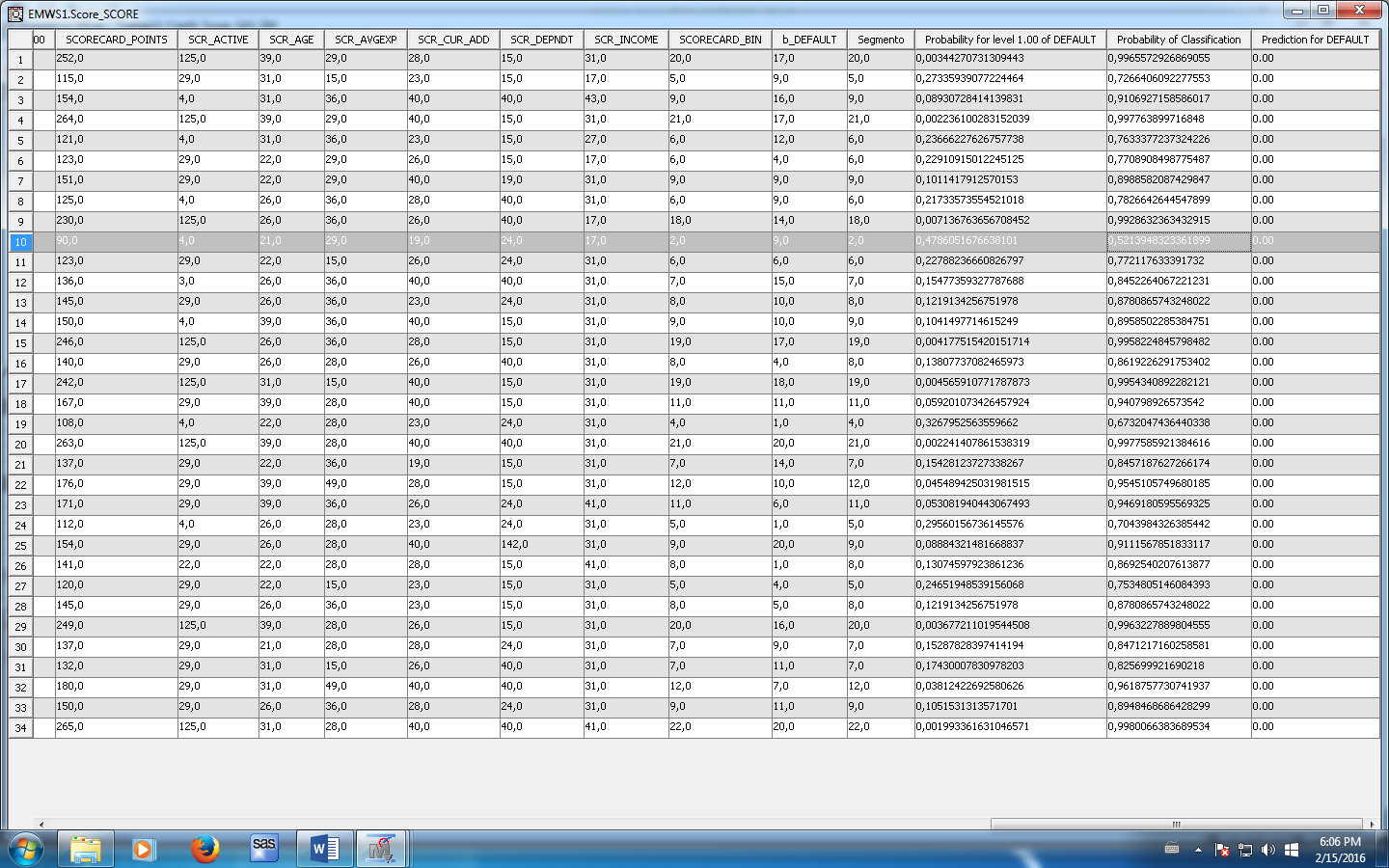
Abajo se puede ver todo el flujo completo para generar la tabla de puntuación con SAS para los 34 clientes.



Para ver los resultados de la tabla de puntuación en SAS hay que ir en datos exportados y pinchar en SCORE > Examinar.



La tabla de puntuación generada por SAS para los 34 clientes ha sido:



Abajo se puede ver una tabla con los 34 nuevos clientes.

ID = ID del cliente

% = Es la probabilidad calculada

Aceptado o no el crédito = Es un campo creado por mí para decir si acepto o no.

La regla es dar crédito para cualquier cliente donde él % sea igual o más grande que 85%. La decisión es muy personal mía, en épocas donde hay necesidad de dar crédito a más gente se puede bajar el % y así aceptar más clientes y también tener más riesgos en la operación de dar crédito, también se puede subir para 90% o 95% y así quedar con menos clientes pero con mucho mejor potencial de pagar todo.

|  |  |  |
| --- | --- | --- |
| **ID** | **%** | **Aceptado o no el crédito** |
| 1319 | 0.99800664 | SI |
| 1289 | 0.9977639 | SI |
| 1305 | 0.99775859 | SI |
| 1286 | 0.99655729 | SI |
| 1314 | 0.99632279 | SI |
| 1300 | 0.99582248 | SI |
| 1302 | 0.99543409 | SI |
| 1294 | 0.99286324 | SI |
| 1317 | 0.96187577 | SI |
| 1307 | 0.95451057 | SI |
| 1308 | 0.94691806 | SI |
| 1303 | 0.94079893 | SI |
| 1310 | 0.91115679 | SI |
| 1288 | 0.91069272 | SI |
| 1292 | 0.89885821 | SI |
| 1299 | 0.89585023 | SI |
| 1318 | 0.89484687 | SI |
| 1298 | 0.87808657 | SI |
| 1313 | 0.87808657 | SI |
| 1311 | 0.86925402 | SI |
| 1301 | 0.86192263 | SI |
| 1315 | 0.84712172 | NO |
| 1306 | 0.84571876 | NO |
| 1297 | 0.84522641 | NO |
| 1316 | 0.82569992 | NO |
| 1293 | 0.78266426 | NO |
| 1296 | 0.77211763 | NO |
| 1291 | 0.77089085 | NO |
| 1290 | 0.76333772 | NO |
| 1312 | 0.75348051 | NO |
| 1287 | 0.72664061 | NO |
| 1309 | 0.70439843 | NO |
| 1304 | 0.67320474 | NO |
| 1295 | 0.52139483 | NO |