

## 1. Clasificación del riesgo

Una vez elegido el clúster sobre el que vamos a hacer el modelo, que van a ser las empresas de tamaño pequeño, continuamos con la preparación de datos para el modelo.

### a. Preparación automática de datos

Utilizamos esta herramienta de SPSS Modeler para preparar los datos que entran en el modelo. Dentro de esta herramienta podemos tratar los valores atípicos, los valores nulos y estandarizar las variables. Hemos configurado con las siguientes condiciones:

- Ajuste del tipo de campos numéricos: en datos de entrada y salida, la herramienta determina si hay valores numéricos con un nivel de medición *Ordinal* puede convertirse a *Continuo* o viceversa.
- Sustitución de valores extremos en valores continuos: en datos de entrada, establecemos un corte de 2 desviaciones estándar los valores atípicos, que son sustituidos por el valor de corte.
- Sustitución de valores perdidos por media en valores continuos: sustituimos los valores nulos o perdidos por el valor de la media.
- Sustitución de valores perdidos por el modo en valores nominales: sustituimos los valores nulos o perdidos por el valor del modo.
- Número máximo de valores de campos ordinales: si un campo ordinal tiene mas de diez categorías, se define como continuo.
- Número mínimo de valores de campo continuos: si un campo continuo tiene menos de cinco valores, se redefine como ordinal.
- Escala común: ponemos los campos de entrada en la escala estándar, donde la media es 0 y la desviación estándar es 1.

### Resumen de procesamiento de campos

Campos		N
<u>Destino</u>		1
<u>Predictores</u>		13
<u>Predictores recomendados para su uso en el análisis</u>	Total	7
	Campos originales (no transformados)	1
	Transformaciones de campos originales	6
	Derivados de fechas y horas	0
	Construidos	0
<u>Predictores no utilizados</u>		6

Vemos que tras la aplicación de la herramienta, los datos han sido transformados. Ha habido transformaciones en los valores atípicos pero no ha habido ninguna transformación en valores perdidos:

### Valores atípicos

Tratamiento de valores atípicos	N
Campos continuos para los que se encontraron y recortaron valores atípicos	7
Campos continuos excluidos porque eran constantes tras la gestión de valores atípicos	1

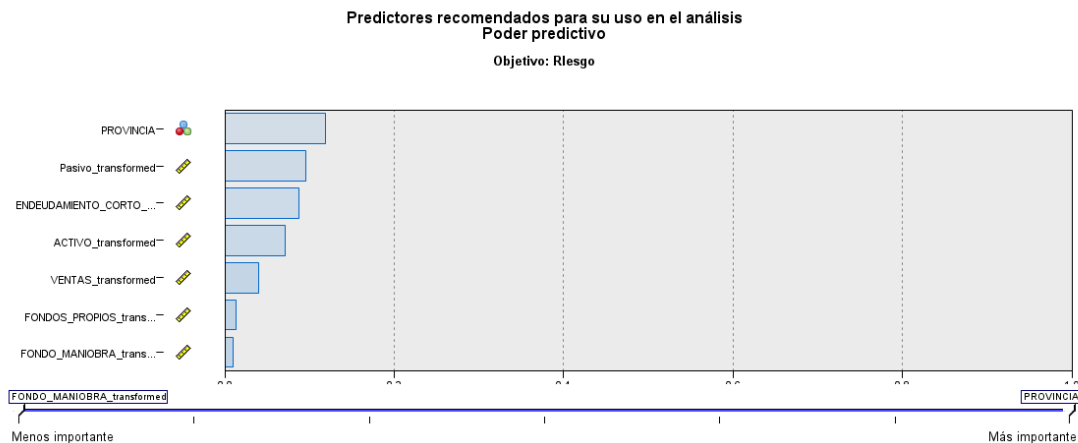
Corte atípico: 2.00 desviaciones estándar por encima o por debajo de la media del campo.

La estandarización se ha llevado a cabo sobre seis predictores:

### Predictores continuos

Transformación	Número de predictores	Criterios	
		Media	SD
Transformar a unidades estándar	6	0,00	1,00

Por último, vemos los predictores que recomienda usar para el modelo:

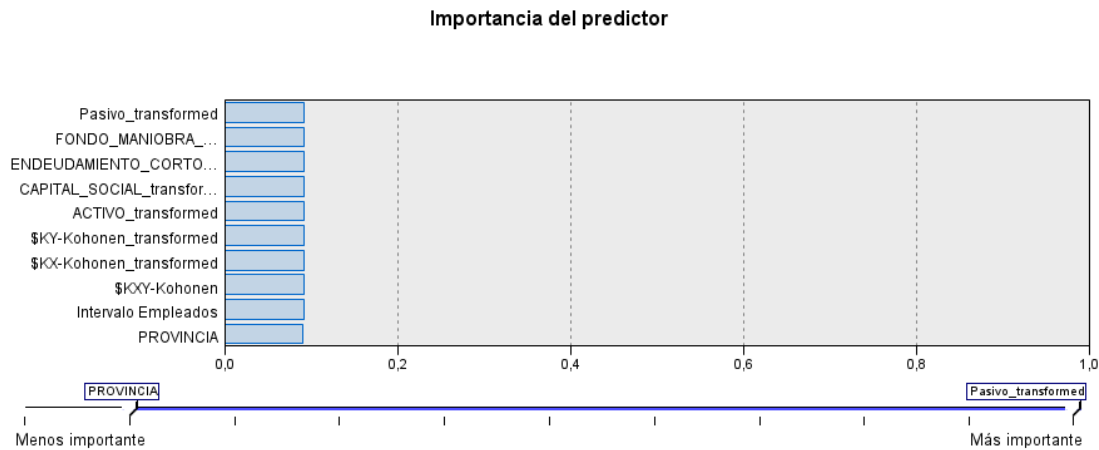


Una vez procesados los datos con esta herramienta, podemos utilizar la siguiente herramienta, que nos ayuda a determinar que modelos funcionan mejor con nuestro conjunto de datos y dependiendo de los objetivos que queramos alcanzar.

#### b. Nodo clasificador automático






Este nodo nos permite especificar el número de modelos a crear, usando los datos de partición, en este caso el entrenamiento, para entrenar el modelo. Seleccionamos todos los modelos posibles para evaluar cuál será la mejor opción para nuestros datos, clasificando modelo en función a la elevación acumulada, que es la tasa de aciertos en cantidades acumuladas con respecto a la muestra global. También vamos a poder evaluar en función a la curva ROC, la precisión global y el beneficio acumulado.

También podemos observar la importancia del predictor calculada, donde muestra un gráfico que indica la importancia relativa de cada predictor al estimar el modelo. Este gráfico está hecho tomando como base todos los modelos seleccionados, por lo que nos da una perspectiva global de las variables importantes.



Los modelos sugeridos

son:

Modelo	Elevación{Sup...	Número de campos	Precisión general (%)	Área debajo de la curva
 Árbol XGBoost 1	2,724	7	94,054	0,870
 XGBoost Lineal 1	1,971	7	93,208	0,719
 CHAID 1	1,808	4	93,256	0,681
 Discriminante 1	1,744	6	80,396	0,678
 Lista de decisiones 1	1,764	3	74,112	0,639

Como podemos ver, los modelos están ordenados por orden descendente de área debajo de la curva, que ha sido la variable principal elegida para la evaluación. Ejecutamos los modelos de árbol XGBoost, XGBoost lineal y CHAID. En todos los modelos, en el apartado de configuración está establecido que riesgo es nuestra variable objetivo.

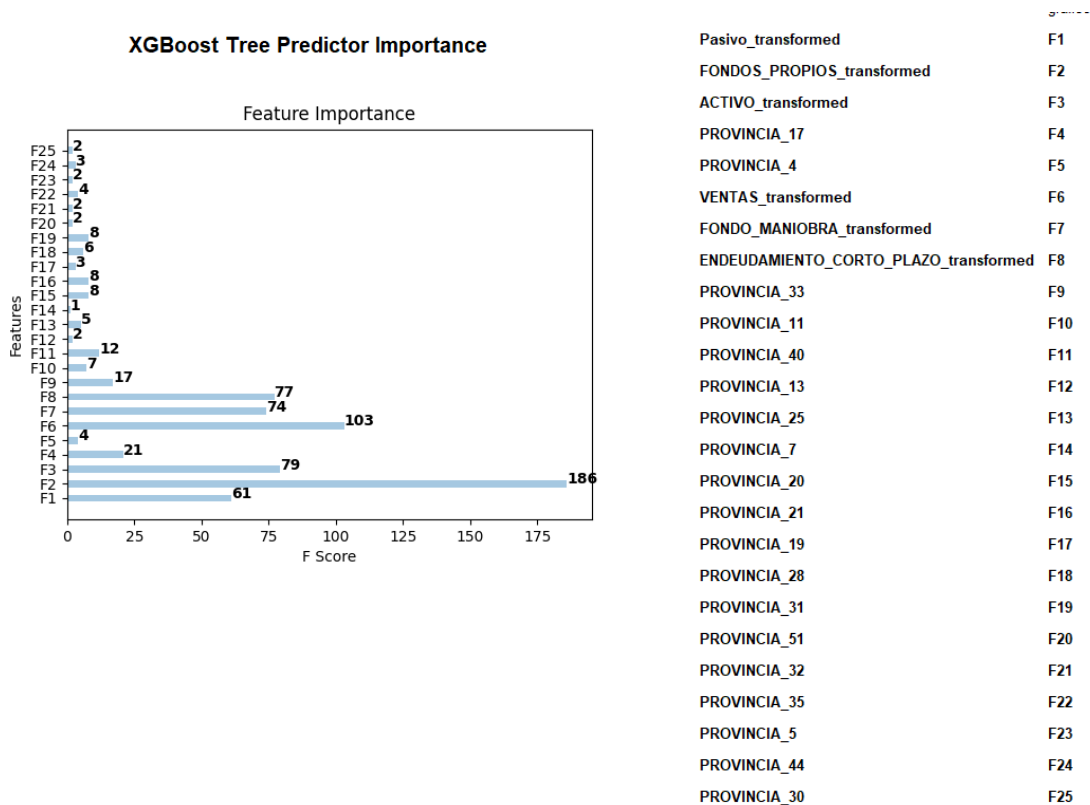
Para evaluar los modelos, hemos comparado los parámetros de área bajo la curva, precisión global y elevación. El área bajo la curva ROC nos proporciona el índice de rendimiento, cuanto más área encontremos bajo la curva, mas exacto será nuestro algoritmo. La precisión nos indica el porcentaje de registros predichos correctamente por el modelo respecto al número total de registros, y la elevación es la tasa de aciertos en cantidades acumuladas respecto a la muestra global, para considerar un modelo válido, debemos obtener por encima de 1,0.

### XGBoost

Es un algoritmo de aprendizaje automático que utiliza el framework de XGBoost (eXtreme Gradient Boosting) para construir un árbol de decisiones a partir de un conjunto de datos. El algoritmo se centra en mejorar la precisión de las predicciones mediante la optimización de los parámetros del modelo y la selección automática de las características más importantes del conjunto de datos.

Con este modelo obtenemos un área bajo la curva de 0.870, con una precisión de 94,054% y una elevación de 2,7.

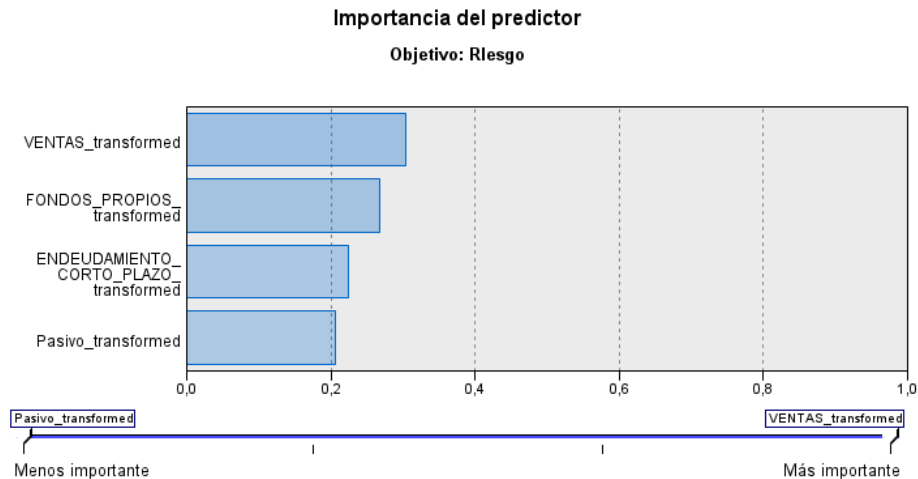
Seleccionamos la opción de optimización de hiperparámetro, que está basada en Rbfopt. Esta opción descubre automáticamente la combinación óptima de parámetros, de manera que el modelo consigue el índice del error previsto o inferior en las muestras.



Como vemos en el gráfico, las variables con más importancia predictiva son los fondo propios, las ventas y el activo.

**CHAID**

Este algoritmo de análisis estadístico se usa para construir árboles de decisión, analiza variables categóricas y encuentra patrones en los datos. Compara la frecuencia de la variable objetivo con las categorías de variables independientes y busca patrones entre las variables para generar un árbol de decisiones jerárquico. Dicho árbol se utiliza para hacer predicciones sobre la variable objetivo en función de las variables independientes.



Con este modelo podemos ver que sus variables predictoras son las ventas de la empresa, fondos propios, endeudamiento a corto plazo y el pasivo.

En este modelo encontramos una precisión del 93,25%, con un área bajo la curva de 0,68 ( el más bajo) y una elevación de 1,8. El árbol tiene una profundidad de 4 nodos y se basa en las siguientes reglas:

```

Pasivo_transformed <= -0,662 [Modo: False] ⇒ False
Pasivo_transformed > -0,662 and Pasivo_transformed <= -0,560 [Modo: False] ⇒ False
Pasivo_transformed > -0,560 and Pasivo_transformed <= 1,286 [Modo: False]
  FONDOS_PROPIOS_transformed <= -0,638 [Modo: False] ⇒ False
  FONDOS_PROPIOS_transformed > -0,638 and FONDOS_PROPIOS_transformed <= -0,411 [Modo: False]
    VENTAS_transformed <= -0,515 [Modo: False]
      FONDOS_PROPIOS_transformed <= -0,465 [Modo: False] ⇒ False
      FONDOS_PROPIOS_transformed > -0,465 [Modo: False] ⇒ False
    VENTAS_transformed > -0,515 [Modo: False] ⇒ False
  FONDOS_PROPIOS_transformed > -0,411 [Modo: False]
    ENDEUDAMIENTO_CORTO_PLAZO_transformed <= 0,310 [Modo: False] ⇒ False
    ENDEUDAMIENTO_CORTO_PLAZO_transformed > 0,310 [Modo: False] ⇒ False
Pasivo_transformed > 1,286 [Modo: False]
  FONDOS_PROPIOS_transformed <= -0,332 [Modo: False] ⇒ False
  FONDOS_PROPIOS_transformed > -0,332 [Modo: False] ⇒ False
  
```

### Algoritmo XGBoost Lineal

El modo lineal en XGBoost utiliza una función de pérdida lineal y una solución de optimización de descenso de gradiente para construir el modelo. En SPSS Modeler, se puede utilizar el nodo XGBoost para ajustar un modelo XGBoost lineal.

En nuestro caso, este modelo nos genera una curva ROC de 0.719 y una precisión general de 93.2%, con una elevación de 1.9.