



***MAESTRÍA EN CIENCIAS EN  
COMPUTACIÓN.***

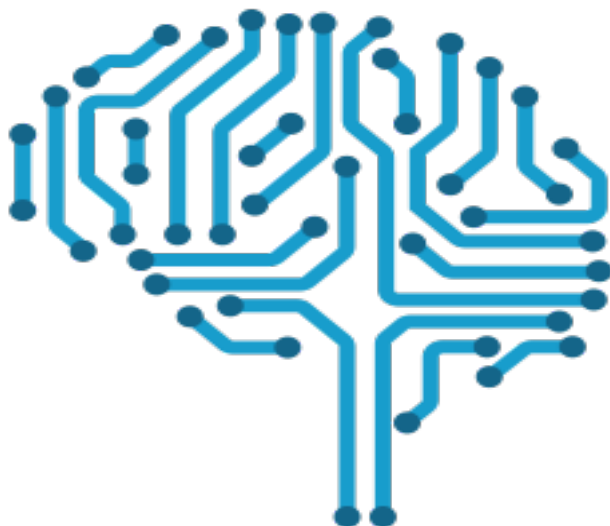
**ARREOLA FRANCO FERNANDO.**

**DR. FERNANDO ESPONDA DARLINGTON.**

**MATERIA: APRENDIZAJE DE MÁQUINA.**

**PROYECTO FINAL.**

**14 DE DICIEMBRE DEL 2017.**



# INTRODUCCIÓN.

Hoy en día, prácticamente cualquier persona puede solicitar un crédito para adquirir bienes y servicios, desde productos básicos hasta propiedades. En algunos bancos o empresas, la flexibilidad para otorgarlo, es demasiada, apenas si se revisa la información proporcionada por el solicitante; en otras, se realiza un análisis exhaustivo del historial crediticio de un determinado cliente y hasta de sus familiares cercanos inclusive, partiremos de la idea de que nos encontramos en este caso. Es esta la razón por la que entra mi motivación para enfocar mi proyecto hacia este tema: La determinación de un crédito es una situación real que día con día, crece y genera millones de solicitudes, de las cuales, hay de todo tipo, como fraudulentas, erróneas, rechazables, aprobables, etc. Las técnicas aprendidas en el curso pueden aplicarse en esta parte, ya que pueden apoyarnos a determinar o predecir la respuesta final que puede ser asignada a una solicitud en curso.

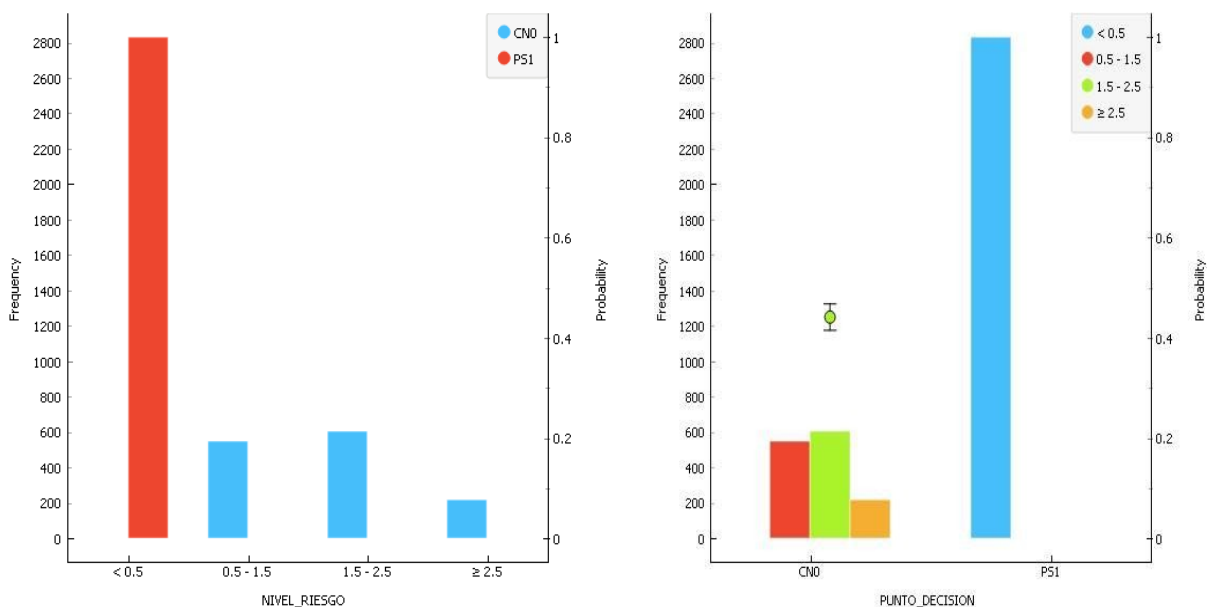
Gracias a la evolución de la computación, el campo del aprendizaje de máquina se ha desarrollado de manera bastante considerable y es ampliamente empleado en la tarea de apoyo en la toma de decisiones para otorgamiento de créditos. Día con día surgen y se mejoran herramientas que pueden facilitarnos y apoyarnos para realizar un determinado análisis, y hay de todo tipo, desde software libre, de licencia, fáciles de emplear, complejas, etc.

El presente proyecto lleva a cabo un análisis básico de un conjunto de datos con información crediticia, siguiendo una metodología establecida y empleando técnicas tales como árboles de decisión, clasificación, naive bayes, regresión logística y reglas de asociación, con el objetivo de interpretar la información que implícitamente está contenida en los datos y poder tener una idea predictiva hacia nuevas solicitudes que se generen. En la segunda parte del proyecto se explica la metodología seguida para la implementación, la tercera parte consiste en interpretar los resultados obtenidos, en la cuarta parte se anexan conclusiones y trabajo futuro y en la última parte se anexa la bibliografía consultada.

# METODOLOGÍA.

Partimos de un dataset que contiene 5201 registros y 64 atributos, el cual, como se mencionó en la introducción, contiene información sobre clientes que están solicitando un crédito. Además del dataset, se cuenta con un catálogo donde se describen los valores que pueden tomar algunos de los atributos, los cuales se describen es el presente reporte cuando lleguen a ser empleados. La metodología seguida consta de los siguientes pasos:

- **HERRAMIENTA A EMPLEAR:** Actualmente se tienen muchas herramientas que implementan técnicas de aprendizaje máquina. En mi caso, me encontraba entre emplear python u orange, comencé con python, pero mi poco conocimiento del lenguaje de programación, estaba haciendo que me tardara mucho en la implementación del proyecto, razón por lo cual, tomé la decisión de emplear la herramienta orange, ya que tengo conocimiento en su uso.
- **ELECCIÓN DE LA VARIABLE OBJETIVO:** Sin tener una variable objetivo, no podemos proceder a realizar análisis alguno, por lo tanto, es el primer paso a seguir. Entendiendo el contexto en el que se desarrolla el proyecto e interpretando los atributos contenidos en el dataset, se llega a dos variables candidatas: PUNTO\_DECISION Y NIVEL\_RIESGO. Para apoyarme en la toma de decisión, decidí graficar ambas variables, obteniendo lo siguiente:



Para poder interpretar lo mostrado por las gráficas, se describen los valores que puede tomar los atributos en cuestión:

#### NIVEL\_RIESGO

- 1 -> Mínimo.
- 2 -> Bajo.
- 3 -> Medio.
- 4 -> Alto.

#### PUNTO\_DECISION

- CN0 -> Decisión de crédito.
- PS1 -> Pre-evaluación de crédito.

Ambas gráficas, en general, nos muestran que para los clientes con riesgo mínimo, se está evaluando si se le otorga un crédito, y para los demás niveles, ya se tiene tomada una determinación. Se empleará como variable objetivo el atributo PUNTO\_DECISION. Justificación: Simplifica y hace más sencillo de interpretar el análisis, ya que tiene una mejor distribución.

- **LIMPIEZA DE DATOS:** Podemos eliminar atributos del dataset, ya que algunos poseen datos irrelevantes o la cantidad de registros es despreciable y no aportan nada para el análisis. Se anexa tabla con los atributos excluidos:

ATRIBUTO	TIPO_DATO	DESCRIPCIÓN	JUSTIFICACIÓN
INFERENCIA_INGRESO_NOMINA	N Numérico	INFERENCIA DE INGRESOS UTILIZANDO COMPROBANTE NOMINA	1 REGISTRO
SUMA_SALDOS_REESTRUCTURA	N Numérico	SUMA DE SALDOS DE CREDITOS REESTRUCTURA	2 REFISTROS
SUMAS_PAGOS_VENCIDOS_REESTRUCTURA	N Numérico	SUMA DE PAGOS VENCIDOS EN CREDITOS REESTRUCTURA	1 REGISTRO
SUMA_PAGO_MIN_REESTRUCTURA	N Numérico	SUMA DE LOS PAGOS MINIMOS EN CREDITOS REESTRUCTURA	1 REGISTRO
SUMA_LINEAS_CREDITO_REESTRUCTURA	N Numérico	SUMA DE LAS LINEAS DE CREDITOS EN REESTRUCTURA	1 REGISTRO
CRED_AUTO_MAS_RECIENTE	Fecha (AAAAMMDD)	FECHA DEL CREDITO AUTOMOTRIZ MAS RECIENTE	2 REGISTROS
MONTO_REFERENCIA_MASANTIGUA	N Entero	MONTO DE SALDO DE LA REFERENCIA MAS ANTIGUA	NO TIENE REGISTROS
PEOR_HISTORIA_TARJETAS	N Entero	PEOR CALIFICACION HISTORICA DE TARJETAS DE CREDITO EN MESES VENCIDOS	NO TIENE REGISTROS
PEOR_HISTORIA_CRED_HIPOTECA	N Entero	PEOR CALIFICACION HISTORICA DE CREDITOS HIPOTECARIOS EN MESES VENCIDOS	1 REGISTRO

Otras modificaciones que se realizaron fueron el cambio de valores en casos donde el registro tenía un valor de 0, se cambia por NULL, y se completaron fechas que estaban incompletas. También se eliminan instancias incompletas, para no meter ruido al análisis.

- **TRANSFORMACIÓN DE DATOS:** Se realiza una discretización para no tener mucha dispersión en la información y facilitar su manejo, con un intervalo de 4, llegando a:

Discretize

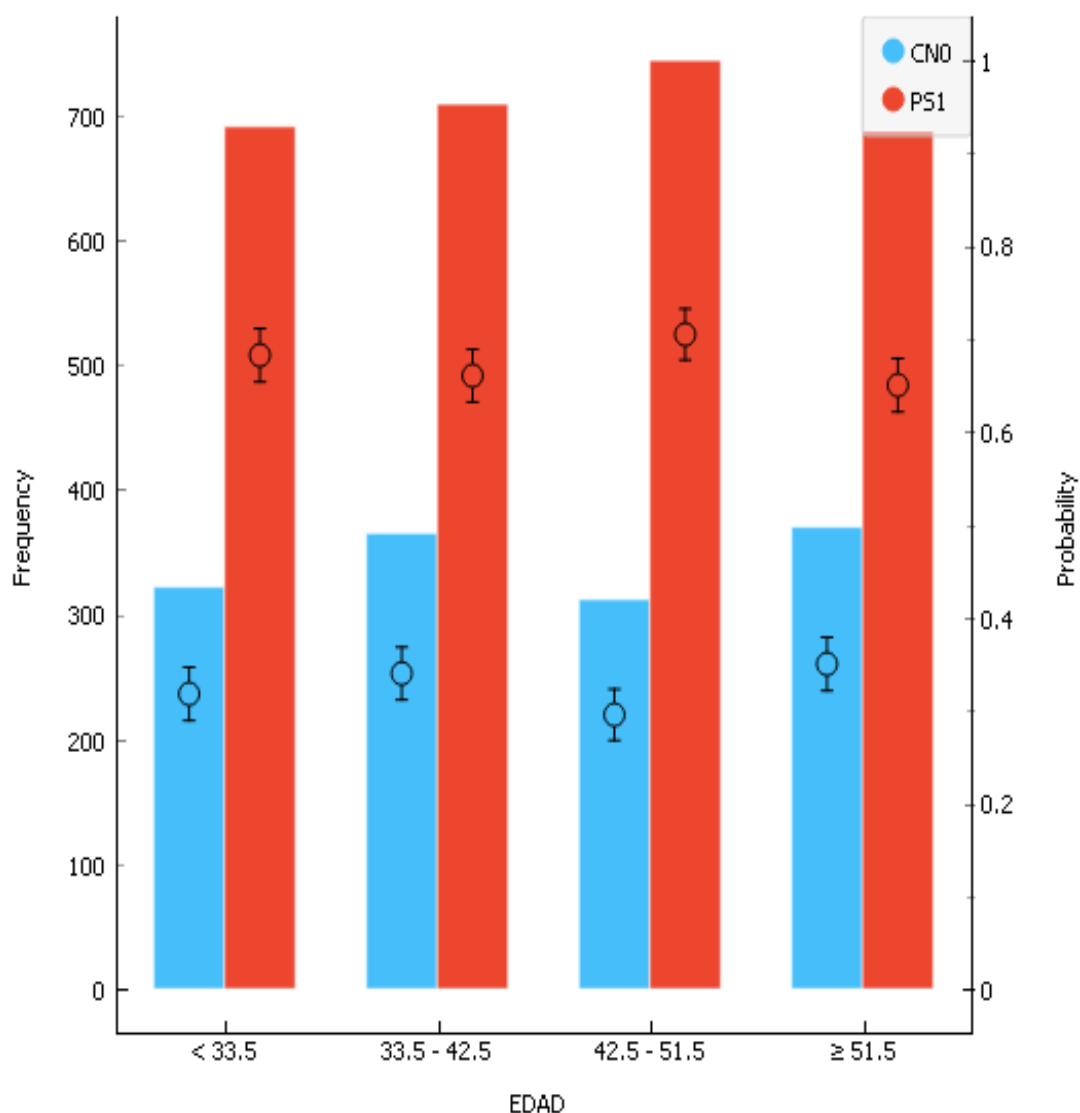
Default method: Equal-frequency discretization

Thresholds

NUM\_SOLICITUD: 20140725001450064.00, 20140805001641300.00, 20140817001845864.00  
PUNTO\_VENTA: 432.00, 4269.00, 7633.00  
SITUACION\_SOLICITUD: 74.00, 75.00  
LINEA\_CREDITO\_FINAL: 0.50, 9849.50, 33499.50  
MESES\_VENCIDOS: 0.50, 1.50, 2.50  
ESTRATEGIA\_LC: 4.00, 8.50, 9.50  
LC\_APLICADA: 0.50, 4.00, 5.50  
INGRESO\_INFERIDO: 658.50, 14336.50, 27423.00  
MODELO\_INGRESO\_INFERIDO: 0.50, 2.50, 4.50  
INGRESO\_FINAL: 750.00, 15009.50, 31529.50  
INGRESO\_DECLARADO: 1.50, 18250.00, 44500.00  
TIPO\_COMPROBANTE\_INGRESO: 10.50, 12.00, 14.50  
INGRESO\_VERIFICADO: 675.00, 10232.00, 29500.00  
MODELO\_INGRESO\_SWA: 2319.00, 7942.00, 11523.00  
MODELO\_INGRESO\_SWB: 1520.50, 6405.00, 9030.00  
MODELO\_INGRESO\_SWC: 2413.00, 8994.50, 14192.00  
MODELO\_INGRESO\_SWD: 1880.50, 8345.50, 10944.50  
MODELO\_NO\_HIT\_INGRESO: 0.50, 1.50, 2.50  
SEGMENTO\_CLIENTE: 0.50, 2.50, 3.50  
SUMA\_LINEAS\_REVOLVENTES\_ONUS: 0.50, 14850.00, 38350.00  
SUMA\_LINEAS\_CRED\_PERSONALES\_ONUS: 0.50, 16499.00, 56279.50  
NIVEL\_RIESGO: 0.50, 1.50, 2.50  
CAPACIDAD\_PAGO\_TOTAL: 282.50, 7149.00, 14996.50  
CAPACIDAD\_PAGO\_NETA: 71.50, 5219.50, 11208.00  
INFERENCIA\_INGRES\_ACTIVOS: 327.50, 28866.50, 57139.50  
SOLICITUDES\_RECHAZADAS: 0.50, 1.50, 3.50  
SCORE\_CLIENTE: 56.50, 213.50, 233.50  
SCORE\_COMPORTAMIENTO: 311.00, 722.00, 740.50  
TARJETA\_MAS\_RECIENTE: 1.50, 1156.50, 1307.50  
COEMITIDA\_MAS\_RECIENTE: 0.50, 1203.50, 1308.50  
PRIVADA\_MAS\_RECIENTE: 52.50, 1104.50, 1304.50  
NUMERO\_TARJETAS\_CRED: 0.50, 1.50, 2.50  
PEOR\_MOP\_TARJETAS\_VIGENTES: 0.50, 1.50, 7.50  
SUMA\_SALDOS\_TARJETAS: 1.50, 3527.50, 12201.50

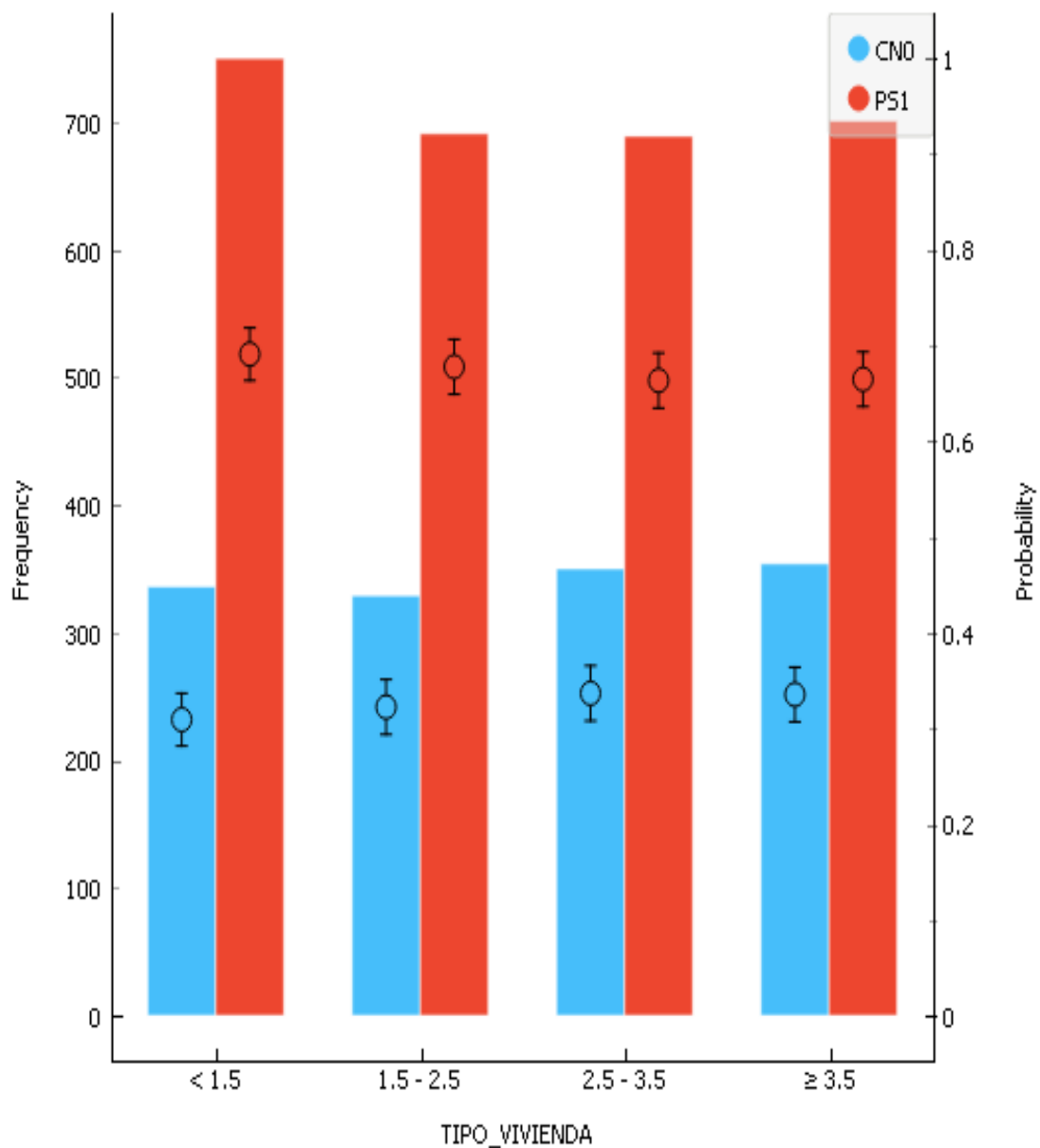
SUMAS\_PAGOS\_VENCIDOS\_TARJETAS: 125.00, 1022.50, 6561.00  
SUMA\_PAGO\_MIN\_TARJETAS: 1.50, 355.00, 945.00  
SUMA\_LINEAS\_CREDITO\_TARJETAS: 0.50, 14450.00, 37850.00  
CRED\_PERSONAL\_MAS\_RECIENTE: 552.50, 1305.50, 1403.50  
FECHA\_CRED\_NOMINA\_MAS\_RECIENTE: 105.50, 1009.50, 1208.50  
FECHA\_CRED\_CHEQUE\_MAS\_RECIENTE: 101.50, 556.00, 707.00  
NUM\_CREDITOS\_PERSONALES: 0.50, 1.50, 2.50  
PEOR\_HISTORIA\_CRED\_PERSONAL: 0.50, 1.50, 36.00  
SUMA\_SALDOS\_CRED\_PERSONAL: 0.50, 18337.50, 56103.50  
SUMAS\_PAGOS\_VENCIDOS\_CRED\_PERSONAL: 1.00, 58.00, 1223.50  
SUMA\_PAGO\_MIN\_CRED\_PERSONAL: 1.00, 784.50, 2773.00  
SUMA\_LINEAS\_CREDITO\_CRED\_PERSONAL: 2500.00, 24000.00, 54500.00  
CRED\_HIPOTECA\_MAS\_RECIENTE: 9960450.50, 19990470.00, 20060929.50  
NUM\_CRED\_HIPOTECAS: 0.50, 1.50  
EDAD: 33.50, 42.50, 51.50  
TIPO\_VIVIENDA: 1.50, 2.50, 3.50  
ESCOLARIDAD: 2.50, 3.50, 4.50

- **ANÁLISIS GRÁFICO DE ATRIBUTOS:** Tomaremos como base cuatro atributos: edad, escolaridad, tipo\_vivienda y situacion\_solicitud, siempre comparados contra nuestra variable objetivo PUNTO\_DECISION. Para la edad, tenemos lo siguiente:



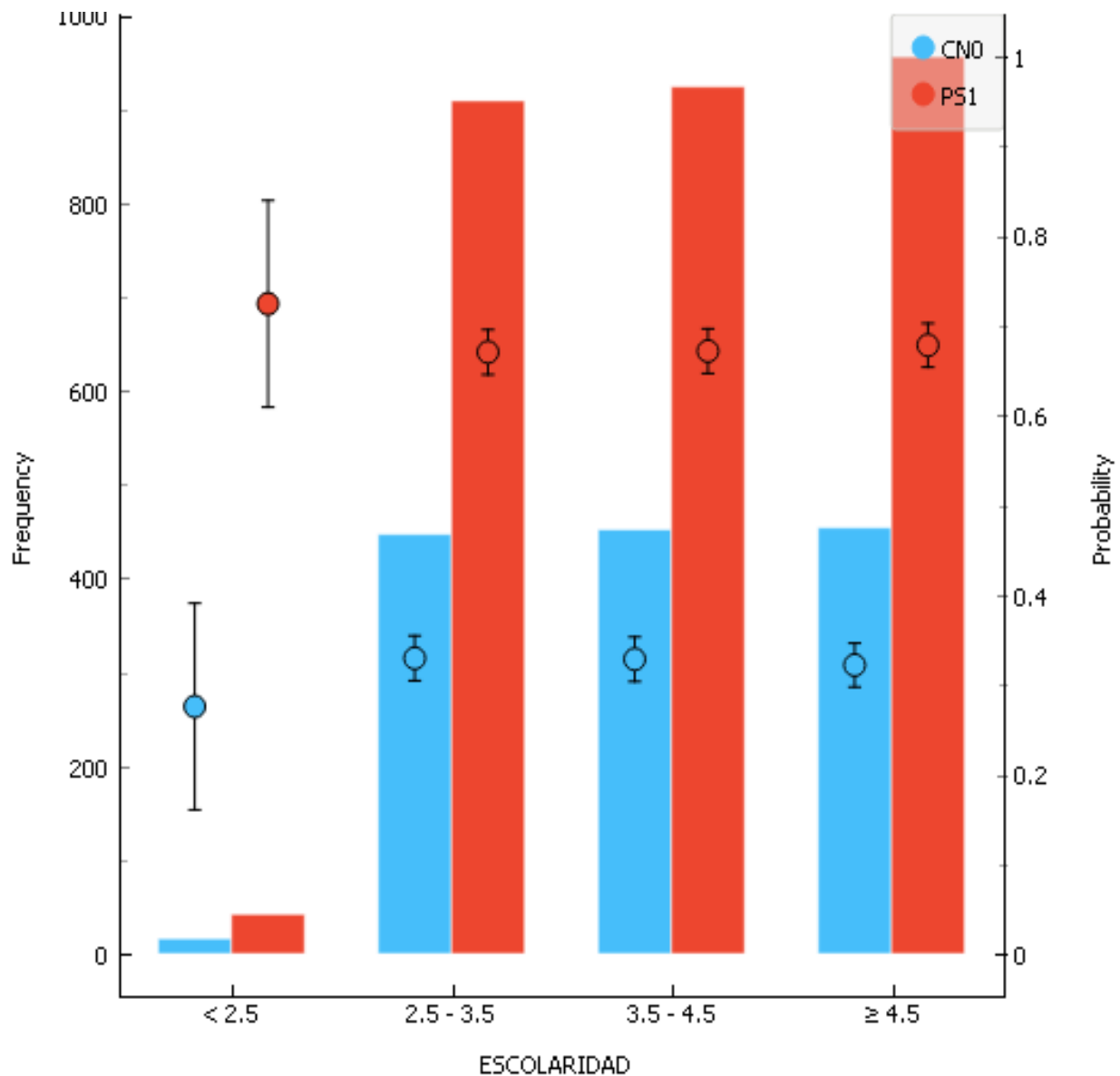
En este caso, la herramienta en automático nos formó los grupos, con base a la discretización mostrada en la tabla anterior. El primer grupo corresponde a personas jóvenes, el segundo a jóvenes-adultos, el tercero a personas maduras y el cuarto a personas mayores. Interpretando las gráficas, podemos notar que en el grupo jóvenes-adultos y personas mayores se tiene el mayor porcentaje de decisión de crédito; podemos pensar que son las edades en las que más y menos créditos se dan. En el caso de la pre-evaluación de crédito, se tiene un porcentaje mayor en el caso de las personas maduras, lo cual se debe a que en promedio a esta edad, bueno o malo, ya se cuenta con cierto historial crediticio que sin duda afecta el tomar una determinación.

Para tipo\_vivienda: Los valores correspondientes son 1 para vivienda propia, 2 para vivienda rentada, 3 para hipoteca y 4 cuando es vivienda con familiares.



Las distribuciones de datos siguen siendo con base a la discretización y los valores de las tablas anteriores. Podemos apreciar que en los casos de vivienda hipotecada y con familiares se tiene un porcentaje mayor de valores correspondientes a decisiones de crédito, que puede traducirse a que este tipo de clientes en más riesgoso otorgarles uno. Caso contrario a las personas con vivienda propia, que tienen un valor de pre evaluación de crédito alta y es muy probable que se les otorgue un crédito.

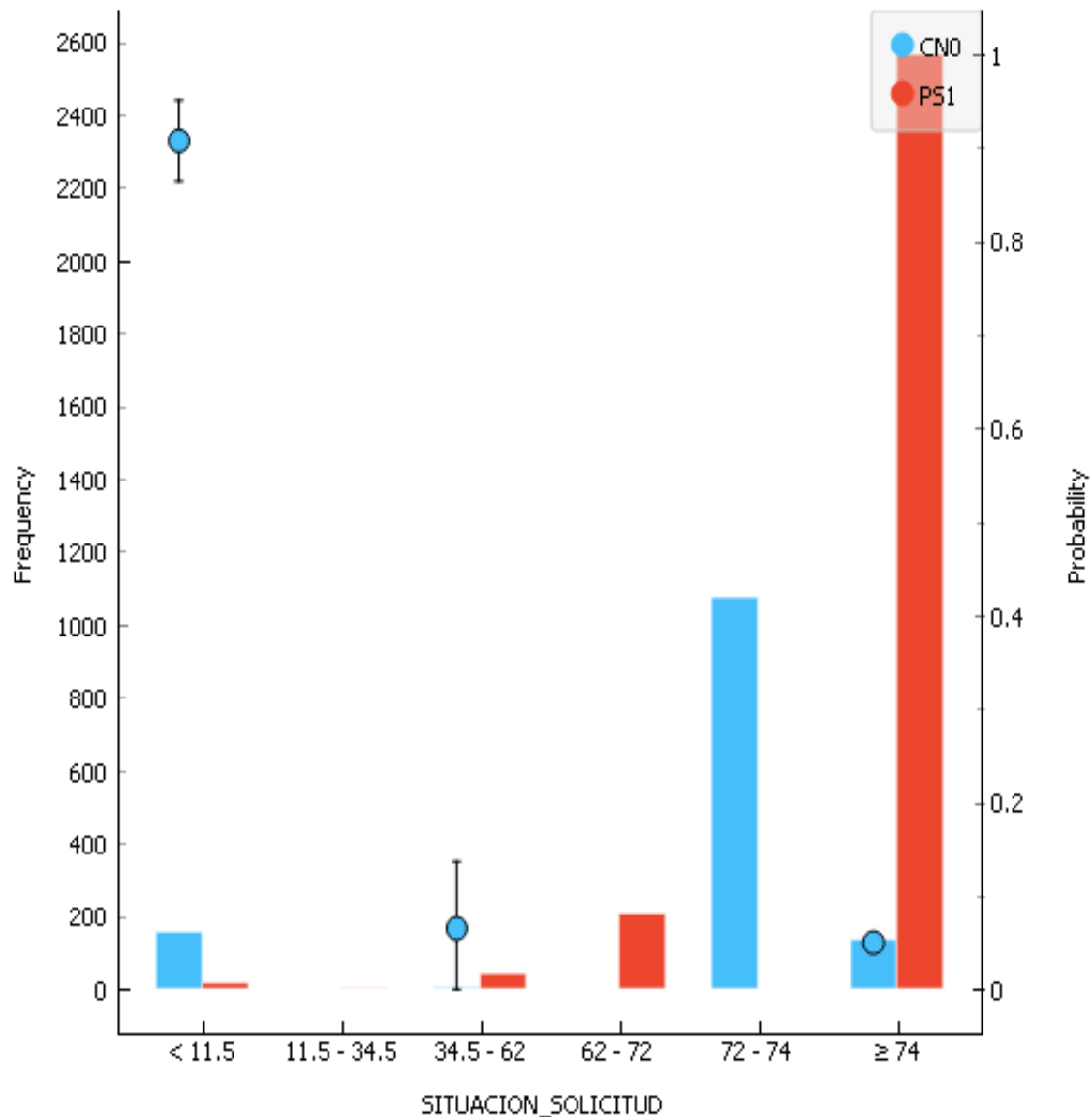
Para escolaridad: Se tiene un valor de 1 para primaria, 2 para secundaria, 3 para preparatoria, 4 para licenciatura y 5 para posgrado.



Se tiene un sesgo muy marcado con las persona que sólo tienen el nivel primaria y secundaria, no hay muchas solicitudes de ese grupo y se sigue teniendo como base los 4 intervalos señalados en la discretización. Los demás grupos no están tan dispersos, respecto a decisión de crédito están a la par, sólo el grupo de gente con posgrado tiene una ligera preferencia respecto a una pre evaluación de crédito y puede interpretarse que se tiene mayor probabilidad de otorgárles una respuesta afirmativa.



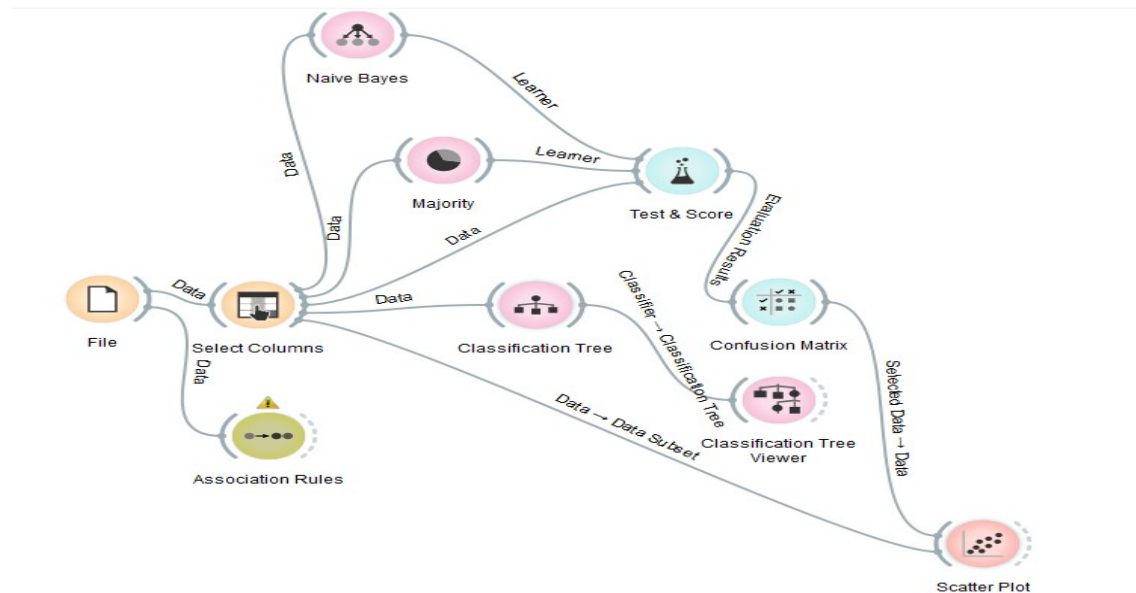
Para situacion\_solicitud: El valor 7 corresponde a una solicitud en proceso, 16 a una solicitud con error, 53 a una solicitud cancelada, 71 a una solicitud pendiente, 73 a una solicitud aceptada y 75 a una solicitud rechazada. Cabe mencionar que el status de rechazada, es el que se toma por defecto.



Podemos apreciar distintas cosas: No se tienen registradas solicitudes con error, muy pocas solicitudes fueron canceladas y las mismas se encontraban por buen camino, se tiene un buen porcentaje de aprobaciones y por último, la mayoría de las solicitudes se estaban evaluando a la hora de extraer los datos, por eso se tienen muchos casos de pre evaluaciones de crédito con el status de rechazada, no todas van a quedar en esa situación y como puede verse, algunas están completamente declinadas.

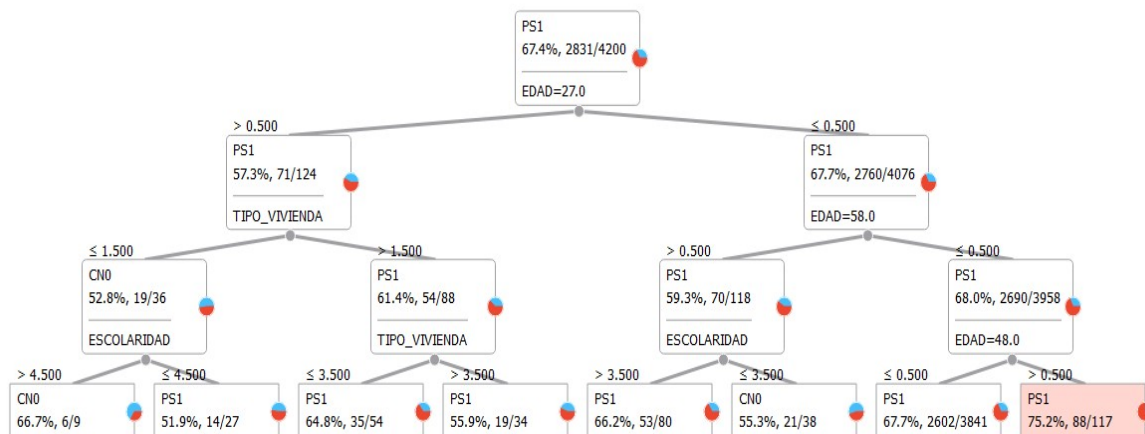
# RESULTADOS.

Esta tercer sección del proyecto consiste en aplicar técnicas de aprendizaje de máquina a nuestro problema en cuestión. Se listan las distintas técnicas aplicadas. Se parte del siguiente diagrama elaborado en la herramienta orange:

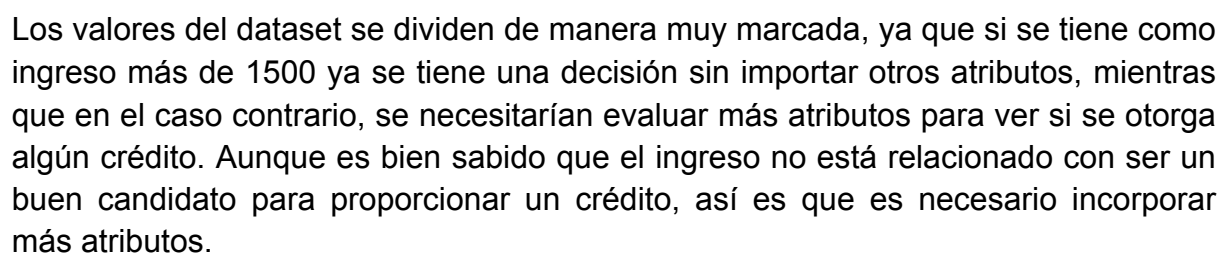


## ➤ ÁRBOLES DE DECISIÓN.

Caso 1: Tomando como objetivo el atributo punto\_decision y como entradas las variables edad, escolaridad y tipo\_vivienda, se llega a:



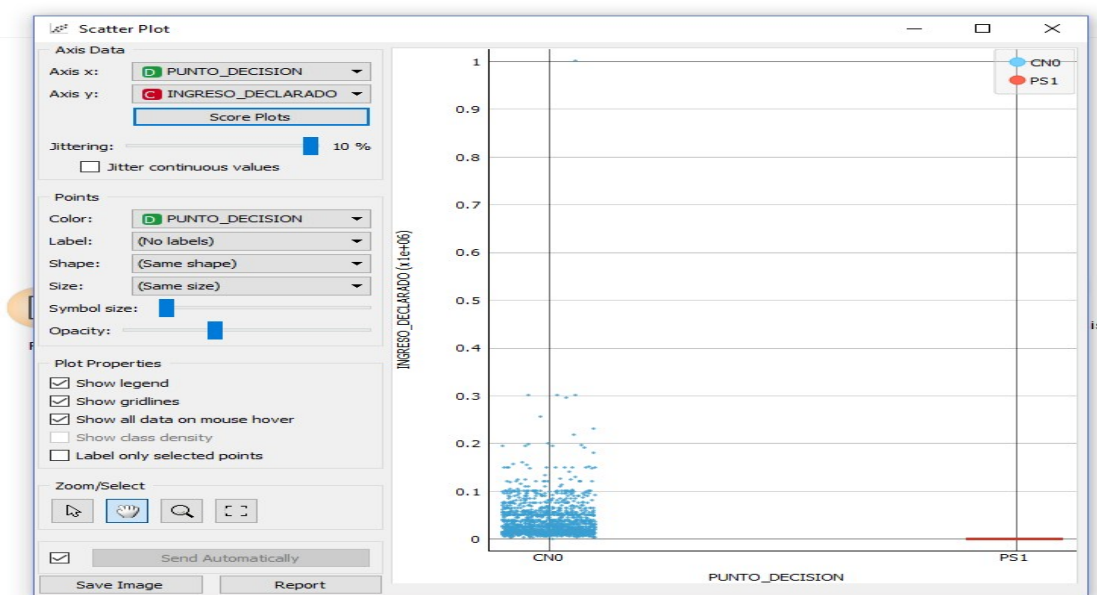
Caso 2: Tomando como objetivo punto\_decision y como entrada ingreso, se llega a:

[illegible]

Se obtiene que el número de solicitudes de las cuales no se ha tomado una decisión de crédito es mayor que el número de solicitudes que ya se ha tomado una decisión, teniendo porcentajes de 0.674% contra 0.326% respectivamente, números que dan a entender que hay que implementar mejoras en el proceso de dar un crédito.

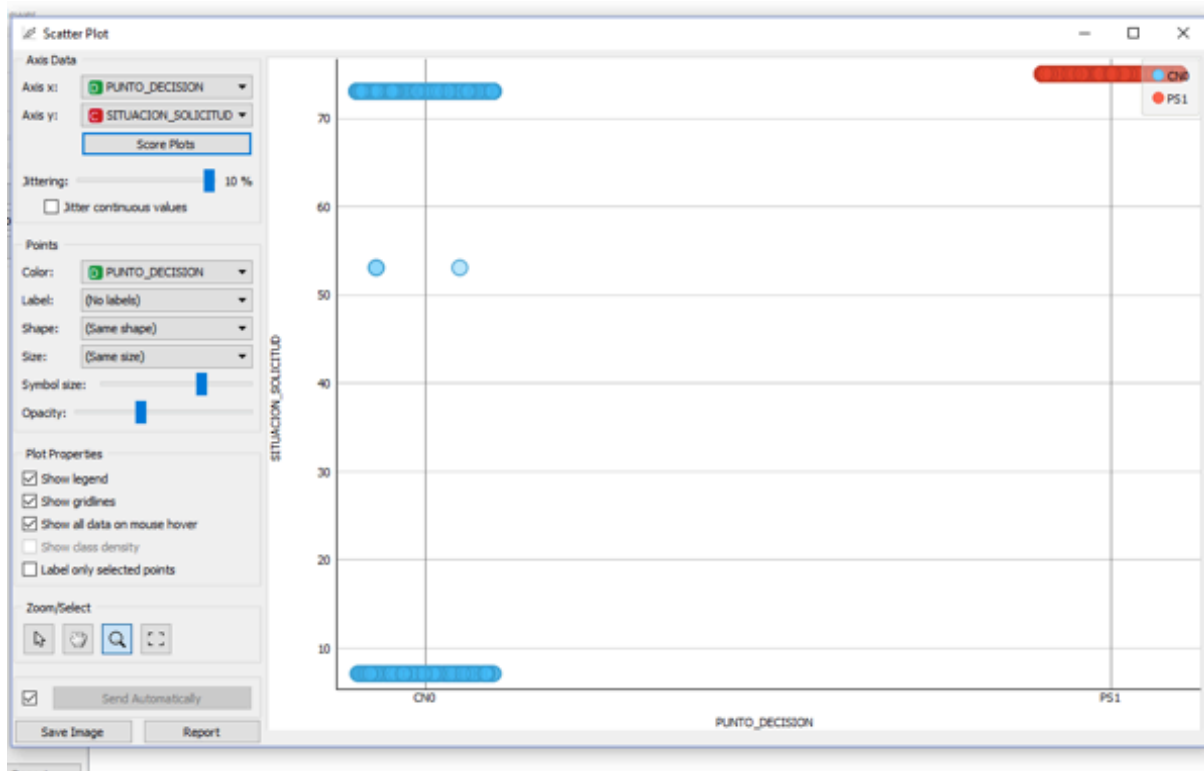
- **NAIVE BAYES Y MATRIZ DE CONFUSIÓN:** Se aplica esta técnica con nuestra variable objetivo contra diferentes atributos, obteniendo las siguientes distribuciones de datos:

Respecto a ingreso\_declarado:



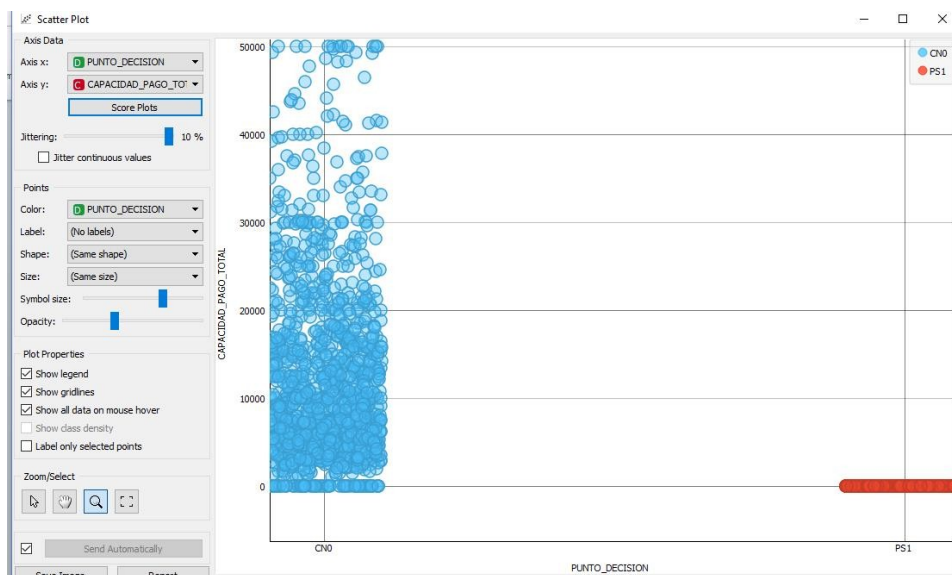
Quedan bien definidos dos grupos, donde hay un gran contraste, ya que en la pre evaluación de crédito se tiene un valor de cero, lo cual puede deberse a que el solicitante no proporcionó sus ingresos o a que hubo algún error en la captura de datos.

Respecto a situacion\_solicitud:



Podemos visualizar la concentración de la información en tres grupos principales, que corresponden a solicitudes en proceso, solicitud aprobada y solicitud rechazada, aunque recordemos que es el status por defecto.

Respecto a capacidad\_de\_pago\_total:



Una representación bastante similar a la arrojada por ingreso\_declarado, aunque la distribución no es la misma, ya que los valores contenidos entre ingreso\_declarado

y capacidad\_de\_pago\_total no son los mismos, ya que el atributo que estamos analizando, depende de otras cuestiones, como si el cliente tiene crédito hipotecario, automotriz, etc.

- **REGLAS DE ASOCIACIÓN:** Se establece una relación entre edad, tipo\_vivienda y punto\_decision, y se obtiene lo siguiente:

Association Rules

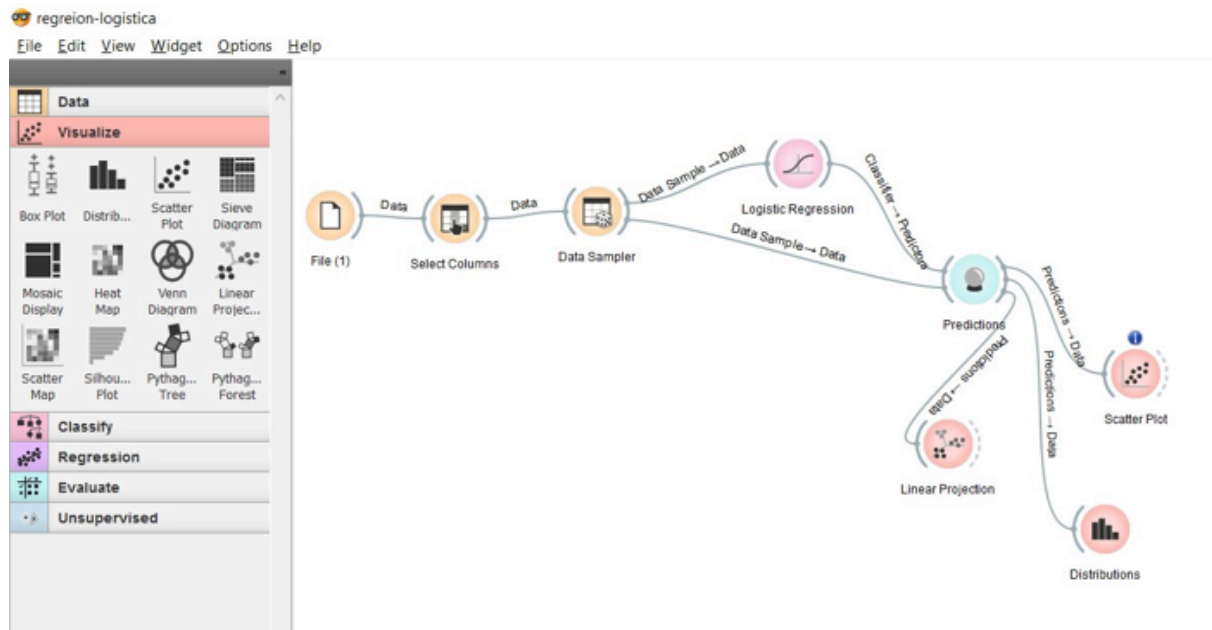
⚠ Data has continuous attributes which will be skipped.

Info	Supp	Conf	Covr	Strg	Lift	Levr	Antecedent	Consequent
Number of rules: 110	0.01	0.83	0.01	80.89	1.23	0.00	EDAD=34.0, TIPO_VIVIENDA=1.0 →	PUNTO_DECISION=PS1
Filtered rules: 85	0.01	0.76	0.01	97.62	1.13	0.00	EDAD=34.0, TIPO_VIVIENDA=2.0 →	PUNTO_DECISION=PS1
Selected rules: 0	0.01	0.70	0.01	94.37	1.04	0.00	EDAD=34.0, TIPO_VIVIENDA=4.0 →	PUNTO_DECISION=PS1
Selected examples: 0	0.01	0.71	0.01	67.40	1.06	0.00	EDAD=49.0, TIPO_VIVIENDA=1.0 →	PUNTO_DECISION=PS1
	0.01	0.90	0.01	91.32	1.34	0.00	EDAD=49.0, TIPO_VIVIENDA=2.0 →	PUNTO_DECISION=PS1
	0.01	0.76	0.01	74.50	1.13	0.00	EDAD=29.0, TIPO_VIVIENDA=3.0 →	PUNTO_DECISION=PS1
	0.01	0.72	0.01	97.62	1.07	0.00	EDAD=29.0, TIPO_VIVIENDA=4.0 →	PUNTO_DECISION=PS1
	0.01	0.76	0.01	74.50	1.13	0.00	EDAD=38.0, TIPO_VIVIENDA=4.0 →	PUNTO_DECISION=PS1
	0.01	0.73	0.01	76.51	1.08	0.00	EDAD=52.0, TIPO_VIVIENDA=1.0 →	PUNTO_DECISION=PS1
	0.01	0.74	0.01	83.26	1.09	0.00	EDAD=52.0, TIPO_VIVIENDA=3.0 →	PUNTO_DECISION=PS1
	0.01	0.68	0.01	74.50	1.02	0.00	EDAD=36.0, TIPO_VIVIENDA=1.0 →	PUNTO_DECISION=PS1
	0.01	0.68	0.01	83.26	1.00	0.00	EDAD=36.0, TIPO_VIVIENDA=4.0 →	PUNTO_DECISION=PS1
	0.01	0.71	0.01	83.26	1.05	0.00	EDAD=57.0, TIPO_VIVIENDA=2.0 →	PUNTO_DECISION=PS1
	0.01	0.83	0.01	78.64	1.24	0.00	EDAD=57.0, TIPO_VIVIENDA=4.0 →	PUNTO_DECISION=PS1
	0.01	0.69	0.01	80.89	1.02	0.00	EDAD=41.0, TIPO_VIVIENDA=3.0 →	PUNTO_DECISION=PS1
	0.01	0.76	0.01	69.05	1.12	0.00	EDAD=32.0, TIPO_VIVIENDA=1.0 →	PUNTO_DECISION=PS1
	0.01	0.72	0.01	97.62	1.07	0.00	EDAD=50.0, TIPO_VIVIENDA=3.0 →	PUNTO_DECISION=PS1
	0.01	0.76	0.01	76.51	1.12	0.00	EDAD=50.0, TIPO_VIVIENDA=4.0 →	PUNTO_DECISION=PS1
	0.01	0.78	0.01	88.47	1.16	0.00	EDAD=45.0, TIPO_VIVIENDA=2.0 →	PUNTO_DECISION=PS1
	0.01	0.79	0.01	97.62	1.18	0.00	EDAD=45.0, TIPO_VIVIENDA=3.0 →	PUNTO_DECISION=PS1
	0.01	0.67	0.01	85.79	0.99	-0.00	EDAD=45.0, TIPO_VIVIENDA=4.0 →	PUNTO_DECISION=PS1
	0.01	0.67	0.01	78.64	0.99	-0.00	EDAD=59.0, TIPO_VIVIENDA=2.0 →	PUNTO_DECISION=PS1
	0.01	0.69	0.01	88.47	1.02	0.00	EDAD=59.0, TIPO_VIVIENDA=4.0 →	PUNTO_DECISION=PS1
	0.01	0.76	0.01	97.62	1.13	0.00	EDAD=37.0, TIPO_VIVIENDA=1.0 →	PUNTO_DECISION=PS1
	0.01	0.72	0.01	88.47	1.07	0.00	EDAD=58.0, TIPO_VIVIENDA=1.0 →	PUNTO_DECISION=PS1
	0.01	0.72	0.01	97.62	1.07	0.00	EDAD=48.0, TIPO_VIVIENDA=2.0 →	PUNTO_DECISION=PS1
	0.01	0.85	0.01	85.79	1.26	0.00	EDAD=48.0, TIPO_VIVIENDA=3.0 →	PUNTO_DECISION=PS1
	0.01	0.72	0.01	78.64	1.07	0.00	EDAD=40.0, TIPO_VIVIENDA=4.0 →	PUNTO_DECISION=PS1
	0.01	0.73	0.01	85.79	1.08	0.00	EDAD=46.0, TIPO_VIVIENDA=2.0 →	PUNTO_DECISION=PS1
	0.01	0.67	0.01	85.79	0.99	-0.00	EDAD=46.0, TIPO_VIVIENDA=3.0 →	PUNTO_DECISION=PS1
	0.01	0.79	0.01	83.26	1.18	0.00	EDAD=47.0, TIPO_VIVIENDA=1.0 →	PUNTO_DECISION=PS1
	0.01	0.68	0.01	91.32	1.01	0.00	EDAD=47.0, TIPO_VIVIENDA=2.0 →	PUNTO_DECISION=PS1
	0.01	0.76	0.01	97.62	1.13	0.00	EDAD=47.0, TIPO_VIVIENDA=4.0 →	PUNTO_DECISION=PS1
	0.01	0.83	0.01	97.62	1.23	0.00	EDAD=51.0, TIPO_VIVIENDA=3.0 →	PUNTO_DECISION=PS1
	0.01	0.81	0.01	104.85	1.21	0.00	EDAD=53.0, TIPO_VIVIENDA=1.0 →	PUNTO_DECISION=PS1
	0.01	0.71	0.01	80.89	1.06	0.00	EDAD=55.0, TIPO_VIVIENDA=1.0 →	PUNTO_DECISION=PS1
	0.01	0.69	0.01	88.47	1.02	0.00	EDAD=60.0, TIPO_VIVIENDA=4.0 →	PUNTO_DECISION=PS1
	0.01	0.70	0.01	85.79	1.03	0.00	EDAD=33.0, TIPO_VIVIENDA=2.0 →	PUNTO_DECISION=PS1

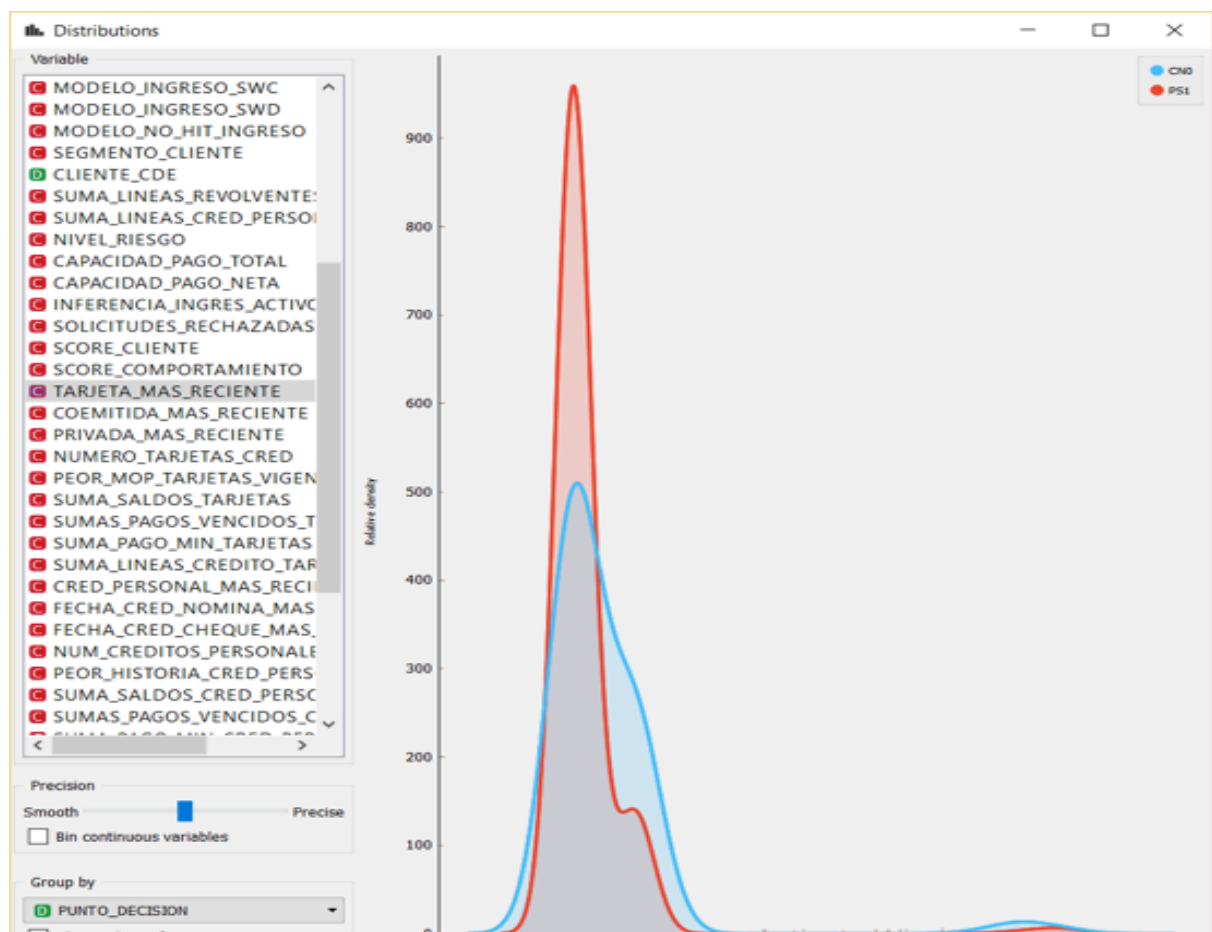
Iniciamente establecí un valor de 0.6 para confianza y 0.5 para soporte, pero no se generó ninguna regla. Ajustando los valores para obtener una buena cantidad de reglas, se llegó a 0.6 en confianza y 0.01 en soporte, un valor que como puede apreciarse ya no es adecuado, ya que ajeno a los valores que tomen edad y tipo\_vivienda, no se llega a una regla confiable.

- **REGRESIÓN LOGÍSTICA:** Se emplea este tipo de regresión debido a que nuestra variable objetivo es categórica, de igual manera, se analizan otras variables contenidas en el dataset.

Partimos del siguiente esquema:



Respecto a tarjeta\_mas\_reciente:

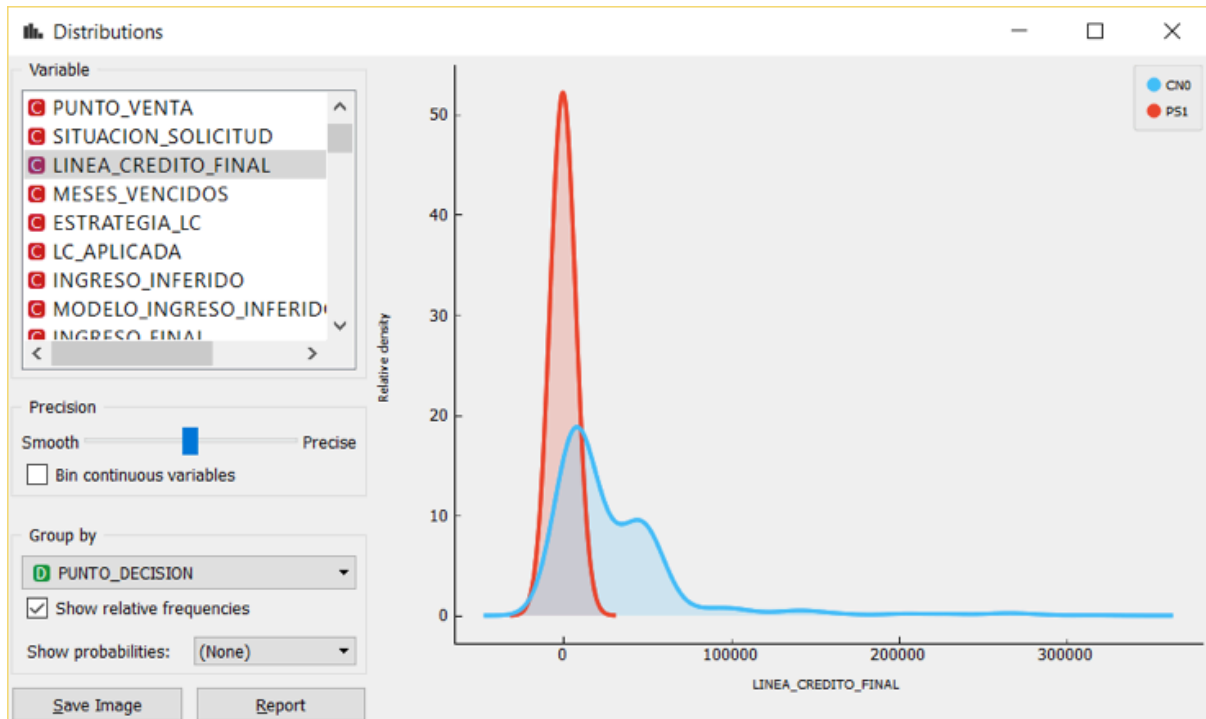


Se aprecian datos interesantes: Las distribuciones mostradas, no están tan lejos de aproximarse a una distribución normal, hecho que facilitaría mucho realizar un análisis más profundo. En el caso de las tarjetas orecientes en el punto de decisión



de pre evaluación, nos representa una información mejor respecto a las otorgadas en una decisión de crédito, debido a que su desviación estándar es menor.

Respecto a linea\_credito\_final:



Hay una mayor concentración de datos en pre evaluación de crédito, con una distribución que a simple vista, parece ser normal, pero sólo a poca gente que se encuentra en este grupo le ha sido otorgado algún crédito, y si compramos los montos, son mucho menores a los proporcionados a las personas que se encuentran en el grupo de decisión de crédito.

## CONCLUSIONES.

Se llevó a cabo con éxito un análisis básico de un conjunto de datos empleando técnicas estudiadas en la materia de aprendizaje de máquina. De manera general y con una comparación entre las técnicas empujadas en la implementación del proyecto, podemos establecer un criterio aceptable para la toma de decisiones sobre otorgar un crédito. Los métodos que aportan más al análisis fueron los árboles de decisión y la regresión logística, las razones quedan justificadas con lo anexado en los resultados.

**Trabajo futuro:** Esperaba que las reglas de asociación proporcionaran una información más confiable, para ello, habrá que realizar un análisis exhaustivo de los datos. De igual forma, limpiar aún más los atributos del dataset, ya que sigue siendo un conjunto grande y seguramente hay más cosas que pueden omitirse. Si bien se



presentó un análisis básico, puede hacerse más robusto y confiable, introduciendo otras técnicas, como redes neuronales, k-means y máquinas de soporte vectorial, para tener un modelo predictivo completo y altamente confiable.

## **BIBLIOGRAFÍA.**

- 1) Orange documentation.
- 2) Esponda F. "Apuntes de clase, aprendizaje de máquina". ITAM, 2017.
- 3) Hastie T., Tibshirani R., Friedman J. "The elements of statistical learning". Second edition, Springer, 2008.