ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE SISTEMAS

PROGRAMA DE MAESTRÍA EN SISTEMAS DE INFORMACIÓN, MENCION INTELIGENCIA DE NEGOCIOS Y ANALÍTICA DE DATOS MASIVOS

"DISEÑO DE UN MODELO DE CREDIT SCORING QUE FOMENTE LA INCLUSIÓN SOCIAL AL SISTEMA FINANCIERO HACIENDO ÉNFASIS EN LAS CARACTERÍSTICAS DEL BIEN A FINANCIAR MEDIANTE TÉCNICAS DE INTELIGENCIA ARTIFICIAL EXPLICABLES"

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL GRADO DE MAGISTER EN SISTEMAS DE INFORMACION

ANDRÉS JULIÁN OQUENDO VILLAMIZAR
andres.oquendo@epn.edu.ec

Director: Dr. Marco Eduardo Molina Bustamante
marco.molinab@epn.edu.ec

Codirector: MSc. Carlos Estalesmit Montenegro Armas
carlos.montenegro@epn.edu.ec

2021

APROBACIÓN DEL DIRECTOR

Como director del trabajo de titulación “DISEÑO DE UN MODELO DE CREDIT SCORING QUE FOMENTE LA INCLUSIÓN SOCIAL AL SISTEMA FINANCIERO HACIENDO ÉNFASIS EN LAS CARACTERÍSTICAS DEL BIEN A FINANCIAR MEDIANTE TÉCNICAS DE INTELIGENCIA ARTIFICIAL EXPLICABLES” desarrollado por Andrés Julián Oquendo Villamizar, estudiante del Programa de Maestría es Sistemas de Información, mención Inteligencia de Negocios y Analítica de Datos Masivos, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la Defensa oral.

Dr. Marco Eduardo Molina Bustamante

DIRECTOR

DECLARACIÓN DE AUTORÍA

Yo, Andrés Julián Oquendo Villamizar, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Andrés Julián Oquendo Villamizar

DEDICATORIA

Dedico este trabajo a mis hijos, como muestra de dedicación y constancia.

No importan los años, cualquier etapa de la vida es buena para estudiar; no importan los éxitos que se haya tenido, el reto de crecer académicamente supera a esos triunfos; no importan los fracasos, siempre podremos continuar si decidimos mejorar.

AGRADECIMIENTO

Agradezco a mi esposa por creer que lo podía hacer y por su insistencia para que lo haga.

Agradezco a mis profesores, pues hicieron que valga la pena.

Agradezco a mis tutores que me guiaron en este trabajo.

ÍNDICE DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE DEFINICIONES

LISTA DE ANEXOS

RESUMEN

Actualmente, en el Ecuador, las instituciones financieras, a pesar de la gran cantidad de datos transaccionales existentes para predecir la probabilidad de pago, continúan utilizando para este fin técnicas tradicionales basadas en modelos estadísticos que se centran básicamente en el análisis de la capacidad de pago y este enfoque por un lado no fomenta la inclusión social al sistema financiero y por otro alienta un comercio ilegal de datos personales. Esta masificación de los datos, que además son propios de cada institución financiera, requiere para su tratamiento de técnicas de Inteligencia Artificial que aprovechan de mejor manera las capacidades computacionales actuales.

En esta tesis se pretende diseñar un modelo de *Credit Scoring* que brinde una mejor solución para predecir el pago de un microcrédito automotriz añadiendo a la solución variables tales como las características del bien financiado, destino, actividad económica del prospecto y el beneficio que esta adquisición con lleva.

En este trabajo se demuestra la efectividad de las técnicas de Inteligencia Artificial, sobre todo de aquellas no denominadas “cajas negras”, cuando adicionalmente se utilizan datos que no están relacionados con la capacidad financiera de los prospectos. Para esto se planteó una metodología para la adquisición y cuantificación de la información de los prospectos en el proceso de calificación y el posterior tratamiento de los datos.

En este proyecto se diseñó un artefacto que mejora el proceso de calificación de los prospectos a un crédito automotriz. Al final de este trabajo se muestra un Sistema de Información creado a partir de los datos resultantes del despliegue de este modelo, el análisis de esta visualización por parte del departamento de mercadeo podría mejorar la efectividad en el otorgamiento de los créditos ya que les ayudaría en el enfoque de las campañas durante la promoción de los créditos.

Palabras clave: Modelo de puntuación para créditos automotrices. Inclusión financiera. Protección de datos personales.

ABSTRACT

Nowadays, in Ecuador, to predict payment probabilities, financial institutions continue to use traditional techniques based on statistical models that analyze the capacity of an individual to pay, this even though there is a huge amount of transactional data that could better predict the probability of repayment and encourage financial inclusion, while at the same time preventing the illegal trade of personal data. This collection of data, which is unique to each financial institution, requires Artificial Intelligence techniques for its handling to better take advantage of the available computational capacities.

This master thesis designs a Credit Scoring model that offers a better solution to predict the repayment of automotive microcredits, adding to the solution variables such as the characteristics of financed goods, destination of funds, economic activity of the candidate, and the benefits that the acquisition will bring.

This work shows the effectiveness of Artificial Intelligence techniques, especially those not called “black boxes”, when additionally, data that is not related to the financial capacity of prospects are used. To achieve this, it proposes that an acquisition and quantification methodology be applied to the prospect’s information during the qualification process and later, during the handling of data.

In this project, an artifact that improves the scoring process of applicants of automotive microcredits was designed. At the end of this work an Information System created from the data resulting from the deployment of this model is shown, the analysis of this visualization by the marketing department could improve the effectiveness in the granting of credits since it would help them in the focus of the campaigns during the promotion of credits.

Keywords: Automotive' Credit Scoring model. Financial inclusion. Protection of personal data.

1. Introducción

En Ecuador el sector automotriz comenzó en el siglo XX con los primeros importadores y distribuidores de vehículos. En la década de los 50's la producción automotriz empieza con la fabricación de carrocerías, asientos, partes automotrices y algunas piezas metálicas. En la década de los 60's se empezaron a fabricar diferentes elementos necesarios para los nuevos modelos de la época, este desarrollo fue favorecido por la Ley de Fomento Productiva. Posteriormente, el modelo ISI (Modelos de Industrialización por Sustitución de Importaciones) facilitó el desarrollo de la industria puesto que buscaba remplazar los bienes importados por bienes producidos localmente, lo que dio paso al nacimiento y desarrollo de las ensambladoras, mismas que por más de tres décadas han fabricado vehículos en el país.

La creación del Plan de Vehículo Popular incrementó la producción automotriz en un 54,21% en el año 1988, es decir, pasó de la producción de 7.864 vehículos en 1987 a 12.127 en el posterior año (https://repositorio.uasb.edu.ec/bitstream/10644/3060/1/T1120-MFGR-Simba%C3%B1a-Desarrollo.pdf pg.12). En 1993, Ecuador empieza a exportar los vehículos ensamblados internamente y se apertura las importaciones hacia Colombia y Venezuela. Adicionalmente, el Convenio Automotor y la dolarización de Ecuador en el año 1999 permitieron que este sector se recupere después de la crisis que afrontó en los 90's y lograra atender la demanda represada de los anteriores años.

En los últimos doce años, las marcas de vehículos más vendidas en el país son Chevrolet, Kia, Hyundai, Great Wall y Toyota. El mercado lo lidera Chevrolet que solo del 2008 al 2017 ha producido y vendido localmente 550.558 unidades, que representa el 59,89 % de la venta total de las cinco marcas. La segunda marca más vendida en el mercado es Hyundai, seguida por Kia, Toyota y Great Wall, destacando de la última una diferencia importante con la primera.

En la Tabla 1, se puede observar las ventas anuales de las distintas marcas desde el año 2007 hasta el 2018.

No se muestran los datos correspondientes al año 2019 y 2020 porque durante la pandemia hubo cambios muy significativos y nuestro propósito es analizar la situación del sector automotriz en condiciones normales.

Tabla 1: Venta de Vehículos por Marca (Top 5) para los últimos 10 años

Año	CHEVROLET	KIA	HYUNDAI	GREAT WALL	TOYOTA
2007	36.174	2.867	9.951	8	7.848
2008	47.519	4.149	13.167	36	10.360
2009	40.185	5.432	11.814	19	6.372
2010	53.429	10.908	17.241	679	8.722
2011	59.189	11.965	14.879	2.085	6.730
2012	54.947	10.144	12.296	2.088	6.840
2013	50.195	12.300	9.629	1.688	6.425
2014	53.574	12.038	10.623	2.160	6.476
2015	40.265	7.647	5.678	2.445	3.651
2016	28.375	8.486	4.930	2.717	2.951
2017	41.101	18.223	9.443	6.792	4.804
2018	45.605	23.141	13.568	8.380	7.947
Totales	550.558	127.300	133.219	29.097	79.126

Fuente AEADE - Anuario 2018

En la Figura 1, se puede observar los valores de la Tabla 1 en un gráfico, en el cual se distingue de mejor manera la significativa diferencia de la marca Chevrolet respecto al resto de marcas.

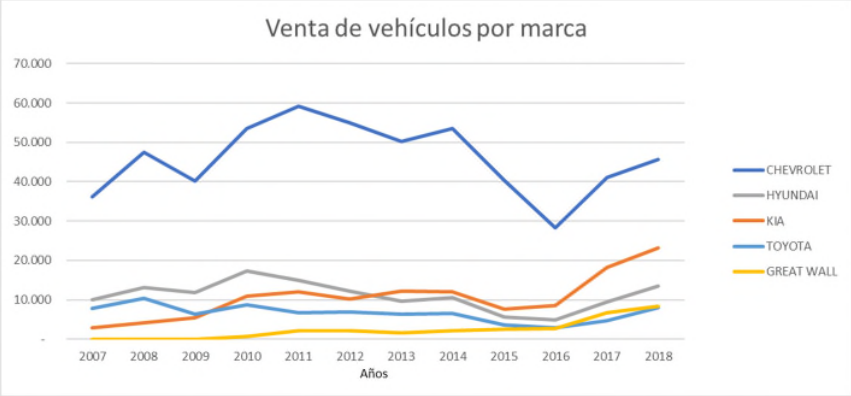


Figura 1: Venta de Vehículos por Marca (Top 5) para los últimos 10 años

Fuente: AEADE - Anuario 2018

La demanda de vehículos en el país se abastece con la producción local y la importación de unidades. En la Tabla 2 se observa el número de unidades exportadas e importadas, además de las ventas de producción nacional y de vehículos importados.

Tabla 2: Resumen de la industria automotriz por unidades

Año	Exportación	Importación	Ventas de producción nacional	Ventas de vehículos importados	Ventas totales
2008	22.774	70.322	46.782	65.902	112.684
2009	13.844	40.649	43.077	49.687	92.764
2010	19.736	79.685	55.683	76.489	132.172
2011	20.450	75.101	62.053	77.840	139.893
2012	24.815	66.652	56.395	65.051	121.446
2013	7.211	62.595	55.509	58.303	113.812
2014	8.368	57.093	60.273	59.784	120.057
2015	3.274	33.640	44.210	37.099	81.309
2016	716	31.761	31.738	371.817	403.555
2017	640	70.203	40.138	64.939	105.077
2018	1.595	101.416	36.818	100.797	137.615

Fuente: AEADE - Anuario 2018

En el último año de la Tabla 2, es decir 2018, la importación y exportación de vehículos ha aumentado, sin embargo, su diferencia es considerable, siendo la importación solo el 1,57% aproximadamente de las exportaciones realizadas durante este periodo de tiempo. Además, las ventas de la producción nacional son alrededor de la tercera parte de las ventas de vehículos importados.

En la Figura 2 se observa que en los últimos años el número de vehículos importados tiende a un crecimiento lineal, mientras que las unidades ensambladas localmente disminuyen considerablemente.

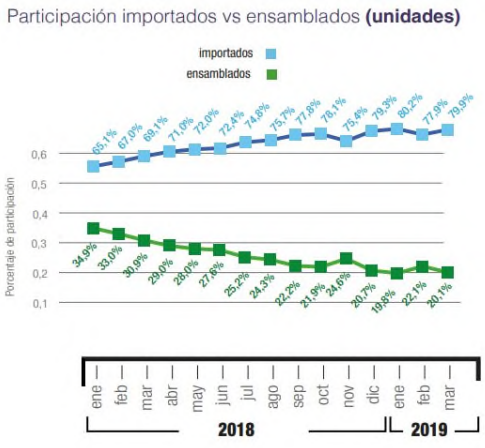


Figura 2: Importación y ensamblaje

Fuente: AEADE - Anuario 2018

Finalmente, en la Tabla 3 se muestra el registro de la importación de vehículos en unidades por segmento, donde se evidencia una mayor importación de automóviles que de cualquier otro segmento.

Tabla 3: Importación de vehículos por segmento en unidades

Año	Automóviles	SUV	Camionetas	Camiones	VAN	Buses
2008	32.585	13.569	9.038	12.654	1.915	561
2009	15.709	12.779	5.343	5.327	919	572
2010	38.418	15.807	13.964	7.390	2.938	1.168
2011	32.090	15.088	9.782	11.148	5.264	1.729
2012	27.545	12.908	10.064	11.930	2.962	1.513
2013	20.099	14.945	6.292	14.281	5.082	1.896
2014	18.820	14.530	5.292	12.615	5.367	469
2015	10.846	8.107	2.948	8.716	2.672	351
2016	13.145	8.213	2.242	4.378	2.358	1.425
2017	30.956	22.698	6.304	5.046	3.455	1.744
2018	44.218	35.079	8.026	7.766	4.886	1.441

Fuente: AEADE - Anuario 2018

La economía ecuatoriana es influenciada en gran medida por el sector automotriz, puesto que crea fuentes de empleo durante los procesos de producción y de comercialización. De hecho, el sector automotriz involucra a varias ramas de actividad económica como la metalmecánica, la petroquímica, la textil, la transferencia tecnológica y otras involucradas

en la fabricación y comercialización de partes, piezas y maquinaria para la producción de este sector. Por tanto, cualquier variación en la producción de automotores, afectará a la economía del país.

De igual forma, la cadena de distribución automotriz involucra a varios actores. En la Figura 3, se muestran los distintos actores, tales como los concesionarios, distribuidores, subdistribuidores en los cuales también participan las financieras.

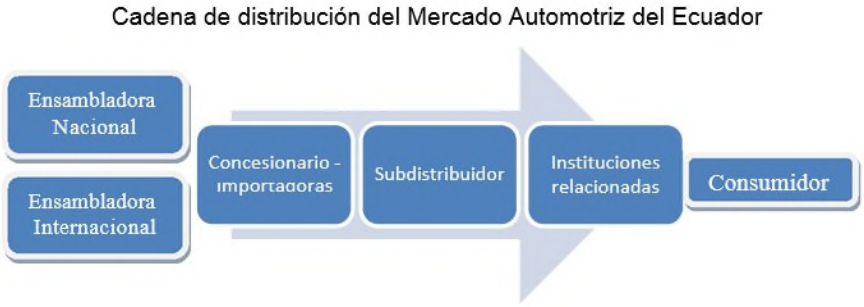


Figura 3: Mercado Automotriz en Ecuador

Fuente: AEADE - Anuario 2018

El crecimiento de la producción nacional e importación automotriz originó el desarrollo de instituciones relacionadas con este sector como los órganos reguladores, el Consejo Nacional de Tránsito, las Cámaras de Comercio e Industria, Instituciones del sector financiero como Financieras y Aseguradoras y otras que intervienen antes y después de la compra - venta de un automotor. En realidad, la comercialización de los vehículos ensamblados en el país y de los importados se realiza a través de los concesionarios autorizados, que sirven como barrera comercial entre las ensambladoras y los consumidores.

Los concesionarios autorizados se encargan de impulsar la venta de vehículos, incentivar la demanda a través de estrategias de marketing, intervenir en el precio del automotor, fomentar la libre competencia, absorber las variaciones de la demanda, brindar servicio postventa, aseguramiento y financiamiento, y en gran medida procurar el crédito automotriz. Los centros autorizados de comercialización consideran las posibilidades del consumidor para adquirir un automotor y sopesan el otorgamiento de créditos directos o a través de una institución financiera.

En el año 2018 el parque automotriz registra 2.056.213 vehículos, de los cuales en los últimos diez años este creció en un 113%, donde el incremento anual es de aproximadamente 130.000 vehículos al año. En la práctica, el mayor número de ventas de vehículos se realizan en las ciudades de Quito y Guayaquil. Un poco más de la mitad del

parque automotor se concentra en las provincias de Pichincha y Guayas, quienes captan el 28 % y 25,8 %, de esas ventas, respectivamente. Azuay es la siguiente provincia en captación, sin embargo, está muy distante de las dos primeras, ya que únicamente tiene el 6,5 % de participación en la distribución.

El incremento anual de vehículos está relacionado con la falta de transportación masiva eficiente y de calidad, lo cual sucinta un mayor consumo de vehículos privados e individuales para la movilización. Según estudios realizados por la Secretaría de Movilidad del Distrito Metropolitano de Quito en la mayoría de los vehículos que circulan en la ciudad se traslada una solo persona.

Composición del mercado automotor ecuatoriano

Durante el 2018 el sector automotor creció un 31 % en relación con el 2017, y cerró el mercado con 137. 615 unidades vendidas, cifra similar a la registrada en el 2011. Esto ocurrió gracias a una mejora de la economía, la expansión del crédito y la eliminación de una serie de restricciones que limitaban la comercialización de vehículos nuevos en Ecuador, lo cual propició un cambio en el mercado automotriz ecuatoriano.

La demanda represada de años anteriores en conjunto con las condiciones de la economía nacional, la política comercial de mayor apertura y el apoyo de las instituciones financieras a través del crédito forjó la recuperación de este sector y la mejora de la economía. Esta demanda represada se originó con el establecimiento de cupos de importación y salvaguardias que impuso el Gobierno, y tuvo su fin con la entrada en vigor del Acuerdo Comercial entre Ecuador y la Unión Europea el 1 de enero de 2017. Parte de los compromisos asumidos en ese proceso de negociación fue que Ecuador cumpliera sus obligaciones internacionales en el marco de la Organización Mundial de Comercio (OMC).

La supresión de cupos restrictivos a las importaciones de vehículos repercutió en el ingreso de los vehículos, lo que fomentó la evaluación de la cartera de productos de las empresas para adaptarse a la demanda del consumidor del mercado ecuatoriano. Entre los años 2016 y 2018, se intensificó la participación de los automotores de origen chino, mexicano, colombiano y europeo en el mercado ecuatoriano originando la política de una apertura comercial. A pesar de ello, los vehículos ensamblados en Ecuador y de origen coreano continúan ocupando los dos primeros lugares en participación. Por otro lado, durante este periodo de tiempo el rango económico de las ventas de vehículos comprendidas en el segmento de un máximo USD 20.000 se acrecentó en 4,1 puntos porcentuales, de 37,1% a 41,2%.

La participación de los diferentes segmentos en el mercado está liderada por los SUV (Vehículo Utilitario Deportivo), cuya participación pasó del 26,8 % en 2016 al 32,8 % en

2018. En términos de ventas, este segmento alcanzó las 45.139 unidades en 2018 puesto que su popularidad a lo largo de estos años se ha fortalecido causando que su cuota de mercado crezca en todo el mundo, por ello la mayoría de las empresas ofrecen una alternativa de estas características en su portafolio.

Análogamente, las ventas de vehículos comerciales en 2017 y 2018 destacaron un importante auge para la recuperación del sector productivo, puesto que son herramientas usadas para ciertas actividades económicas como el reparto de carga liviana y pesada, el turismo, movilización educacional, entre otras actividades. Los vehículos comerciales con mayor extensión en el 2018 fueron las VANS con la venta de 4.407 unidades y el crecimiento del 53,8%, seguido por el segmento de camiones con 7.844 unidades vendidas y un incremento del 37,1%, por último, el segmento de buses con ventas de 1.907 unidades y el aumento del 2,9%. Contrariamente, el segmento de las camionetas fue la que obtuvo menor participación en el mercado con ventas de 19.464 unidades durante el año 2018. Ahora bien, se observa que los dos primeros meses del año 2019 se produjo una estabilización de la venta de nuevos vehículos, con aproximadamente 10.000 unidades vendidas.

El crédito automotriz en Ecuador

Los niveles de producción, importación y venta de vehículos, así como el crecimiento automotriz sostenido durante los últimos años se deben a factores como la dolarización, el envío de remesas de los migrantes y el financiamiento para la adquisición de estos bienes. Estos tres factores han permitido mejorar el nivel de vida de los ecuatorianos e incrementar su capacidad de pago, consecuentemente dio apertura a que los consumidores puedan adquirir vehículos de contado, o acceder a los créditos proporcionados por bancos, financieras o empresas orientadas a la compra de cartera.

Pese a ello, existen dos brechas que deben incluirse en el análisis de la economía ecuatoriana en el año 2019. La primera brecha es coyuntural, desde 2017 los créditos entregados por los bancos fueron mayores que los depósitos, lo que propicia la generación de cartera vencida con el paso del tiempo. La segunda es estructural, el fortalecimiento del dólar en conjunto con la política interna que eleva los costos de producción disminuye la competitividad ecuatoriana frente a sus socios comerciales, lo que resta dinamismo a las exportaciones (medidas en volumen) e impulsa las importaciones. Además, la ausencia de otras fuentes permanentes de divisas incita el déficit en la cuenta corriente de la balanza de pagos y provoca la caída sostenida en las reservas internacionales del Banco Central, que solo se recuperan cuando el Gobierno consigue financiamiento en el exterior.

Figura 4: Crecimiento de los depósitos y la cartera de crédito

Fuente: AEADE - Anuario 2018

Como se muestra en la Figura 4, desde el segundo trimestre de 2017 la cartera de créditos de los bancos empezó a mostrar tasas de crecimiento superiores a las de los depósitos. En septiembre de ese año la brecha entre ambas tasas alcanzó su punto máximo: el crecimiento anual de los créditos fue de 23,8 % y el de los depósitos de 7,4 %, es decir, los créditos crecieron a un ritmo más de tres veces mayor. Esa situación fue posible porque en los meses previos, especialmente en 2016, la demanda de créditos de personas y empresas se redujo sensiblemente y los bancos acumularon mucha liquidez (generada en parte por los créditos que el Banco Central entregó al gobierno de entonces). Una vez que la demanda de crédito se reactivó, los bancos empezaron a usar esa liquidez para entregar más préstamos. Si bien la brecha entre el crecimiento de los créditos y el de los depósitos se ha ido cerrando, en enero del 2019 la diferencia seguía siendo de más de ocho puntos porcentuales.

Consecuentemente, el indicador de liquidez de los bancos (fondos disponibles / depósitos de corto plazo) según el Banco Central del Ecuador en el documento “Monitoreo de los principales indicadores monetarios y financieros de la economía ecuatoriana (abril 2019)”, señala que el mencionado indicador llegó a rozar el 35% a finales de 2016, lo que resulta ineficiente tanto para los bancos como para la economía en su conjunto, en marzo del 2019 ya se ubicó en 24,3%. Ese indicador es tres puntos porcentuales menor que el de marzo de 2018 y siete puntos menores que el de 2017, aunque se encuentra en un nivel prudente no es posible mantener ese ritmo de caída por más tiempo. Dado que el ritmo de crecimiento de los depósitos no muestra signos de repunte (su fuerte alza en 2016 e inicios de 2017 respondió sobre todo a la inyección a la economía de los recursos que el Gobierno obtuvo de nueva deuda externa y de préstamos del Banco Central, que ahora están prohibidos), es previsible que el crecimiento de los créditos también se siga desacelerando, afectando al consumo de productos, como los autos, que con frecuencia se pagan con créditos.

Normativa y gestión de riesgo de crédito

Las instituciones financieras ecuatorianas deben considerar, las normas regulatorias vigentes en el Ecuador y los lineamientos de cada institución para mitigar el riesgo al que se enfrenta, y calificar adecuadamente los activos de riesgo y realizar las previsiones adecuadas para resguardar su patrimonio.

Dentro de la normativa vigente, específicamente en el título VII de los Activos y de los límites de crédito se establece que las instituciones financieras deben reflejar la verdadera calidad de los activos y su valoración periódica. Así mismo, es necesario realizar las previsiones para cubrir los riesgos propios de una cartera de crédito, tales como pérdida de valor, incobrabilidad, errores, cambios en las tasas de interés, modificación de leyes y variaciones en el mercado.

La Ley Orgánica de Instituciones del Sistema Financiero contempla el tratamiento del castigo a obligaciones que no han sido recuperadas o estuvieran en mora por tres años. También especifica los límites que tienen las instituciones para realizar operaciones activas y contingentes con personas naturales, jurídicas y grupos financieros. Uno de los aspectos que se relacionan con el crédito y la calificación de las operaciones es la Central de Riesgos, que es un sistema de registro de los deudores de las instituciones financieras que cuenta con información individualizada, consolidada y clasificada que permite conocer el estatus de un cliente.

El Comité de Supervisión Bancaria de Basilea, expone ciertas prácticas y normas que permiten gestionar el riesgo. Una de las normas eficaces es “conocer a su clientela”, esta debe ser aplicada por todas las instituciones financieras y los bancos deberán aplicar las mismas políticas y procedimientos en todas sus filiales y sucursales. El programa consiste en cuatro elementos primordiales: política de aceptación de clientes, identificación de clientes, seguimiento continuo de las cuentas de mayor riesgo y gestión del riesgo.

Dentro del primer principio, de los cuatro básicos del supervisor, el Comité de Basilea establece que los bancos deben evaluar la suficiencia del capital en función al perfil de riesgo y establecer un proceso riguroso que contempla cinco características que son:

- Vigilancia por parte del Directorio y alta Gerencia.
- Evaluación rigurosa del capital.
- Evaluación integral de los riesgos.
- Seguimiento e información
- Evaluación de los controles internos en caso de existir.

La evaluación integral de los riesgos contempla toda clase de exposición de riesgo: operativo, mercado, liquidez, tipo de interés en la cartera de inversión, entre otros. Para

este caso se considerará al riesgo de crédito a los que están expuestos los bancos, financieras y empresas compradoras de cartera automotriz.

De acuerdo con la Ley Orgánica de Instituciones del Sistema Financiero, un crédito de consumo es aquel que se destina al pago de un bien o servicio, no relacionado con la actividad productiva, cuya amortización se realizará a través de cuotas periódicas. Al administrar este tipo de crédito, toda institución financiera debe realizar una adecuada calificación de activos y en este caso, la calificación será por cada operación. Además, debe realizar una adecuada selección de sujetos de crédito, determinar la capacidad de pago y estabilidad de la fuente de recursos.

"La calificación cubrirá la totalidad de la cartera de créditos de consumo concedida por la institución del sistema financiero, según los criterios antes señalados y con base a los parámetros listados en la Tabla 4":

Tabla 4: Parámetros para la concesión de créditos

CATEGORÍAS	DÍAS DE MOROSIDAD
A-1	0
A-2	1 - 8
A-3	9 - 15
B-1	16 - 30
B-2	31 - 45
C-1	46 - 70
C-2	71 - 90
D	91 - 120
E	+ 120

Fuente: Junta de Regulación Monetaria Financiera

Libro I: Sistema Monetario y Financiero

La calificación que cada sujeto de crédito reciba, se basará en: criterios permanentes, la antigüedad de los dividendos no pagados y la totalidad del monto adeudado, compuesto por deudas por vencer, vencidas y de aquellas en las que no se devengan intereses.

1.1 Planteamiento del problema

Las instituciones financieras, a pesar de la inmensa cantidad de datos transaccionales que poseen, continúan utilizando para la puntuación crediticia (Credit Scoring) técnicas tradicionales basadas en modelos estadísticos, que se centran básicamente en el análisis de la capacidad de pago de los prospectos, para lo cual las instituciones financieras en algunos casos incluso obtienen los datos de los prospectos sin su consentimiento y sin seguir las debidas regulaciones, atentando contra su privacidad. Esta caracterización del riesgo, por un lado, impide el acceso al crédito a quienes no tuvieron un buen historial o no han sido parte del sistema financiero, lo cual no permite una mayor inclusión financiera, aspecto que va en contra del crecimiento económico mundial, por añadidura desaprovecha el aporte que podrían dar las características del objeto a financiar, para determinar la probabilidad de pago. En la mayoría de los modelos que actualmente se diseñan se usan datos demográficos y relacionados con la capacidad de pago, datos que generalmente son cuantitativos y suficientes para el diseño de modelos con técnicas tradicionales, como regresión logística, tal como se afirma en [13] y [14], lo que hace innecesario acudir a la Inteligencia Artificial (IA) para la obtención de mejores modelos.

El uso de métodos innovadores podría mejorar la precisión de los modelos para la puntuación crediticia[1]. Sin embargo, su uso también genera preocupaciones sobre la privacidad de los datos, la equidad y el potencial de discriminación contra las minorías, la interpretación de los modelos y el potencial de consecuencias no deseadas porque los modelos desarrollados sobre datos históricos pueden aprender y perpetuar el sesgo histórico**.**

En [1] se incluye siete recomendaciones de políticas de orientación sobre puntuación crediticia, una de las cuales indica que las decisiones tomadas sobre la base de la puntuación crediticia deben ser explicables, transparentes y justas. Este ítem resulta ser un serio problema para las técnicas de caja negra de IA.

Los resultados de los modelos utilizados para la puntuación crediticia deberían ser explicables para que provean una adecuada retroalimentación, lo que facultaría la creación de Sistemas de Información que ayuden a las entidades financieras a descubrir nichos de mercado donde las campañas de mercadeo puedan ser más eficaces.

En este proyecto, se pretende solucionar algunas de las deficiencias de los modelos de *Credit Scoring* que utilizan técnicas estadísticas tradicionales y no utilizan eficientemente datos cualitativos, como los existentes en las características del bien a financiar. También propone el uso de técnicas de IA diferentes a las conocidas como cajas negras, ya que estas no son explicables e impiden el desarrollo de Sistemas de Información.

1.2 Objetivo General

Diseñar un modelo de *Credit Scoring*, para la determinación del riesgo en el otorgamiento de microcréditos automotrices con énfasis en las características del bien financiado para que fomente la inclusión social al sistema financiero, basado en técnicas de Inteligencia Artificial explicables para que en su puesta en producción habilite la creación de un Sistema de Información donde sea posible la visualización y análisis de resultados.

1.3 Objetivos Específicos

- Caracterizar y categorizar el riesgo en el financiamiento automotriz en base al comportamiento del pago de las obligaciones, buscando explicar tal comportamiento, incluyendo las variables relacionadas con el objeto financiado, además de las de estructura financiera y demográfica.
- Descubrir información confiable, para propender al uso de los resultados.
- Recolectar información suficiente con una estrategia de recolección de datos adecuada para evitar sesgos en la técnica.
- Lograr un modelo predictivo que categorice el riesgo de acuerdo con los criterios definidos, que use técnicas de IA y que apoye la gestión crediticia.
- Para los niveles de riesgo crediticio que se definan, caracterizar patrones de comportamiento de riesgo, que sean significativos para cada nivel, de manera que estos resultados apoyen la generación de políticas que permitan el adecuado seguimiento y control de la gestión crediticia, por medio de un Sistema de Información.
- Comparar el comportamiento del modelo de Credit Scoring propuesto, con el modelo obtenido mediante el uso de técnicas estadísticas tradicionales, utilizando datos de prueba obtenidos de la cartera de clientes de un banco del Ecuador.
- Obtener modelos de puntuación crediticia que sean explicables.
- Fomentar el diseño de modelos de puntuación crediticia que fomenten la inclusión Social al sistema financiero.
- Fomentar la adquisición de datos para el diseño de modelos de puntuación crediticia que no atenten contra la privacidad de las personas y no contravengan las leyes.
- Realizar seguimiento y control de los procesos de crédito y cobranza para cada tipo de cliente.
- Gestionar grandes volúmenes de datos dentro del mismo entorno analítico.
- Facilitar el proceso para el establecimiento de límites de riesgo y políticas de crédito.
- Mejorar la calidad de colocación de crédito y por ende la calidad de la cartera.

2. Justificación

2.1 Justificación Teórica

Según la Ley Orgánica de Instituciones del Sistema Financiero[2], la misión fundamental de las instituciones financieras es hacer uso eficiente y seguro de los depósitos de sus clientes y para esto es necesario que las colocaciones de crédito se hagan al grupo adecuado, en el tiempo correcto y con el menor riesgo posible.

Las técnicas de IA, incorporadas en las herramientas de minería de datos, son consideradas como un conjunto de técnicas, algoritmos y herramientas que permiten resolver problemas para los que, a priori, es necesario cierto grado de inteligencia, en el sentido de que son problemas que suponen un desafío incluso para el cerebro humano [3]. Estas técnicas facultan a los científicos de datos, campo que normalmente ha sido copado por estadísticos y matemáticos, a crear modelos de *Credit Scoring*, con la ventaja de que este tipo de modelos matemáticos, según estudios recientes [4], tienen mayor precisión y predicen con mayor agilidad. Por ello, los científicos de datos son capaces de crear modelos personalizados para el sector financiero y muy a pesar de que estos métodos también tienen algunas limitaciones, en general, casi todas las técnicas producen explicaciones fiables y contribuyen a una mayor transparencia de la rendición de cuentas de los sistemas de decisión [5].

Siguiendo la recomendación expresada en [3], relacionada con el presente y futuro de la IA y la cantidad de datos necesarios para utilizar estas técnicas, entendemos entonces que con el volumen de datos transaccionales con que actualmente cuentan las entidades financieras, es posible garantizar el uso adecuado de la IA en el desarrollo de modelos de predicción.

Actualmente, en el Ecuador, las instituciones financieras, a pesar de la gran cantidad de datos transaccionales existentes para predecir la probabilidad de pago y cumplir con su misión, continúan utilizando técnicas tradicionales basadas en modelos estadísticos que se centran básicamente en el análisis de la capacidad de pago [6]–[10], esto en muchos casos, dicho por experiencias propias en trabajos anteriores, la utilización de este tipo de modelos genera controversias: entre los prospectos quienes no entienden el porqué de su puntuación, los vendedores de crédito (mercadeo) que se molestan al ver su trabajo desvanecido y los analistas de crédito que no lo pueden explicar la razón de la puntuación ya que es derivada de una ecuación compleja. El uso de técnicas de Inteligencia Artificial explicables luego de estas controversias mejoró el clima y ayudo a la culminación y otorgamiento del crédito, pues el cliente entendió y respaldo las exigencias de la entidad financiera para asegurarse el pago.

Una vez conocida la finalidad de las instituciones, analizamos a continuación como ha sido el uso de las técnicas matemáticas para este fin.

Para ello revisamos cronológicamente 39 artículos que comparan las técnicas tradicionales con las de IA, desde el 2002 [11] hasta el 2020 [12], lo cual se recoge en la Tabla 5.

Tabla 5: Artículos que comparan técnicas para el diseño de modelos

Tipo	Técnica	Año	Ref.	Título	Autores
Tradicional	Regresión Logística	2005	[13]	Modelling small-business credit scoring by using logistic regression, neural networks and decision trees	M. Bensic, N. Sarlija
		2018	[14]	Credit Risk Prediction Using Artificial Neural Network Algorithm	D. Kumar Gupta y S. Goyal
		2020	[15]	Comparison of Non-Parametric Techniques against Logistic Regression	M. M. Amaro
		2020	[12]	The future of credit scoring modelling using advanced techniques	J. Cermakova
Inteligencia Artificial	Redes Bayesianas	2002	[11]	Learning Bayesian network classifiers for credit scoring using Markov Chain Monte Carlo search	B. Baesens, M. Egmont
	Árboles de Decisión	2006	[16]	Comparing decision trees with logistic regression for credit risk analysis	J. B. Simha
		2008	[17]	Credit scoring with boosted decision trees Credit scoring with boosted decision trees	J. Bastos
		2010	[18]	A comparison of data mining techniques for credit scoring in banking: A managerial perspective	H. Ince y B. Aktan
		2010	[19]	Vertical bagging decision trees model for credit scoring	D. Zhang, X. Zhou, S.
		2012	[20]	Two credit scoring models based on dual strategy ensemble trees	G. Wang, J. Ma, L. Huang
		2014	[21]	A New Credit Scoring Method Based on Rough Sets and Decision Tree	S. E. E. Profile
		2019	[22]	Using decision tree classification algorithm to design and construct the credit rating model for banking customers	F. Shahbazi
		2020	[23]	Machine Learnig in credit scoring	F. Innocenti
		2020	[24]	A comparative study of forecasting Corporate Credit Ratings using Neural Networks, Support Vector Machines, and Decision Trees	P. Golbayani, I. Florescu
	SVM	2007	[25]	Credit scoring with a data mining approach based on support vector machines	C. L. Huang, M. C. Chen
		2011	[26]	Application of Artificial Intelligence Techniques for Credit Risk Evaluation	A. Ghodselahi y A. Amirmadhi
		2020	[27]	Optimized algorithm for credit scoring	A. Chacko, A. Antonidoss
	Redes Neuronales	2011	[28]	Instance sampling in credit scoring: An empirical study of sample size and balancing	S. F. Crone y S. Finlay
		2016	[29]	Customer Credit Risk Assessment using Artificial Neural Networks	N. Mohammadi y M. Zangeneh
		2017	[30]	A Better Comparison Summary of Credit Scoring Classification	S. Imtiaz y A. J
		2019	[31]	Credit scoring to classify consumer loan using machine learning	A. Natasha, D. D. Prastyo
		2019	[32]	How do machine learning and non-traditional data affect credit scoring	L. Gambacorta, Y. Huang
		2020	[33]	Artificial Intelligence in Finance	N. R. Tadapaneni
		2020	[34]	Classification Performance for Credit Scoring using Neural Network	C. Edmond, I. Joumal
	XGBoost	2018	[35]	Performance Evaluation of Machine Learning Approaches for Credit Scoring	A. Cao, H. He, Z. Chen
	Big Data	2018	[36]	Regulatory learning: How to supervise machine learning models? An application to credit scoring	D. Guégan y B. Hassani
		2019	[37]	Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending	B. Niu, J. Ren, y X. Li
	Cuadros de Mando	2011	[38]	Using data mining to improve assessment of credit worthiness via credit scoring models	B. W. Yap, S. H. Ong
		2013	[39]	Credit Scoring Using Machine Learning	K. Kennedy
	Técnicas de Simulación	2017	[40]	Artificial intelligence in engineering risk analytics	D. Wu, D. L. Olson
	Discretized Enriched	2019	[41]	A discretized enriched technique to enhance machine learning performance in credit scoring	R. Saia, S. Carta
	State Machines	2020	[42]	Credit Risk Rating using State Machines and Machine Learning	B. Sabeti, H. A. Firouzjae
	Puntuación de confianza entre pares	2019	[43]	Machine Learning-based Credit Scoring System and Framework o f " Peer Trust Score "	M. R. Kumar
	Combinación de técnicas	2007	[44]	A comparison study of credit scoring models	D. Zhang, H. Huang
		2011	[45]	A comparative assessment of ensemble learning for credit scoring	G. Wang, J. Hao
		2013	[46]	The Use of Genetic Algorithm, Clustering and Feature Selection Techniques in Construction of Decision Tree Models for Credit Scoring	M. Khanbabaei y M. Alborzi
		2016	[47]	Classification methods applied to credit scoring: Systematic review and overall comparison	F. Louzada, A. Ara
		2017	[48]	Credit scoring models: Techniques and issues	Y. L. Eddy, E. Muhammad
		2017	[49]	A Hybrid Machine Learning Approach for Credit Scoring Using PCA and Logistic Regression	S. Walusala, R. Rimiru
		2018	[50]	Application of Ensemble Models in Credit Scoring Models	A. Chopra y P. Bhilare

De la Tabla 5 se desprende el siguiente resumen donde se señala cada técnica y los artículos que la utilizan:

- Regresión Logística: [13][14][15][12]
- Redes Bayesianas: [11]
- Árboles de Decisión: [16][17][18][19][20][21][22][23][24]
- SVM: [25][26][27]
- Redes Neuronales: [28][29][30][31][32][33][34]
- XGBoost: [35]
- Big Data: [36][37]
- Cuadros de Mando: [38][39]
- Técnicas de Simulación: [40]
- Discretized Enriched: [41]
- State Machines: [42]
- Puntuación de confianza entre pares: [43]
- Combinación de técnicas: [44][45][46][47][48][49][50]

Este análisis, refuerza el criterio de que la mejor técnica es aquella que más se ajusta a los datos y, además, que no solo se trata de ajustarse a los datos sino también de su preprocesamiento, como imputación en [30], muestreo en [28], entre otros.

Se concluye de esta comparación, tal como lo sugiere IBM en su herramienta de minería de datos [51], que lo mejor es usar procedimientos que combinen varias técnicas, de tal manera que el resultado final sea una votación de las técnicas con aportes más fiables.

Para nuestro propósito, al diseñar el modelo combinaremos varias técnicas y buscaremos que al menos una de ellas sea un Árbol de Decisión, ya que esta técnica nos permite alcanzar dos objetivos: por un lado, cumple con la directriz referida a que los modelos deben ser explicables y, por otro, incorpora la información descubierta en los Sistemas de Información de las instituciones financieras, enriqueciendo su contenido con Analítica Avanzada (BA), lo cual habilita a los usuarios de estos sistemas, a un acceso unificado al conocimiento embebido en los modelos predictivos. Por ejemplo, un mejor conocimiento sobre los nichos de mercado donde presentar sus ofertas y así aumentar la oportunidad de éxito.

Las complejas técnicas modernas del aprendizaje automático (ML) las cuales son óptimas cuando la cantidad de datos es muy grande, compiten bien con los sistemas de puntuación de crédito tradicionales en términos de potencia predictiva. Sin embargo, la falta de explicabilidad en los resultados dados por los algoritmos de las técnicas llamadas cajas negras crea desconfianza y barreras que finalmente relegan las decisiones de préstamos al consumidor de alto riesgo.

En la comunidad de Inteligencia Artificial, la noción de una “inteligencia artificial explicable” se ha extendido entre sus miembros, cuyo objetivo radica en que los seres

humanos sean capaces de comprender fácilmente los resultados [52], en este contexto los Árboles de Decisión son una técnica que ayuda significativamente en la interpretación de resultados [53].

En el detalle de esta justificación resaltamos algo que apoya nuestro propósito y es que entre todas las técnicas de IA expuestas en los artículos analizados, las más usadas son los Árboles de Decisión, incluso en algunos de estos artículos [17], [19]–[22], [24] sobre esta técnica se realizaron procedimientos adicionales para potenciar su rendimiento y capacidad predictiva.

2.2 Justificación Metodológica

Como se ha descrito anteriormente en este trabajo se busca desarrollar un modelo de *Credit Scoring* con IA, para el financiamiento automotriz con énfasis en las características del bien financiado para que, además de conseguir Inclusión Social en su despliegue y puesta en producción, se pueda crear un Sistema de Información en el cual se visualicen y analicen sus resultados. Es decir, se busca diseñar una solución más efectiva que la dada por las técnicas estadísticas tradicionales. Para la comprensión, ejecución y evaluación de nuestra solución, se decidió utilizar el paradigma de la ciencia del diseño, que caracteriza gran parte de la investigación en la disciplina de Sistemas de Información ya que, según Hevner, en este paradigma el conocimiento y la comprensión del dominio de un problema y su solución se logran con la construcción y aplicación de un artefacto [55].

En el presente trabajo, ese artefacto consiste en un modelo predictivo desarrollado con el uso de métodos de IA, cuyo objetivo es brindar una mejor solución para predecir el pago de un crédito otorgado por una institución financiera y demostrar que las soluciones clásicas, con modelos estadísticos, pueden ser reemplazadas con ventaja, por métodos de IA, es decir, cae dentro de la definición de los objetivos de la ciencia del diseño que indica Hevner: mediante este paradigma la creación de artefactos nuevos e innovadores amplían los límites de las capacidades humanas y organizativas para la resolución de problemas.

Por último, se eligió esta metodología debido a que propone una base matemática para muchos tipos de evaluaciones cuantitativas de un artefacto para un Sistemas de Información, incluidas las pruebas de optimización, la simulación analítica y las comparaciones cuantitativas con diseños alternativos [18]. En el presente trabajo hemos decidido realizar una comparación entre un modelo predictivo construido con IA y otro construido con técnicas estadísticas tradicionales. Adicionalmente, utilizando la misma metodología, se podrían realizar trabajos futuros para la evaluación adicional del modelo

en un contexto organizacional dado, con la posibilidad de aplicar métodos empíricos y cualitativos [18].

2.3 Justificación Práctica

El mercado automotriz en el Ecuador es uno de los sectores que dinamiza la economía, a través de la producción e importación de vehículos y la comercialización de estos. El crecimiento de este sector en los últimos años ha sido constante y sostenido, al igual que el incremento de financiamientos otorgados a los consumidores o sujetos de crédito para la compra de un vehículo.

Adicionalmente, en el Ecuador existe la presión para la promulgación de la ley que proteja los datos personales, sobre la base del Reglamento General de Protección de Datos (RGPD), del 25 de mayo de 2018, promulgado para toda la Unión Europea, cuando esto ocurra y sobre todo, cuando sea aplicable mediante un reglamento, la adquisición de los datos, que típicamente usan los modelos tradicionales, será más compleja y podrá traer consecuencias legales [56]. En tal virtud se hará necesario buscar información adicional que no violente esta ley y los datos correspondientes al objeto a financiar no tendrán carácter personal.

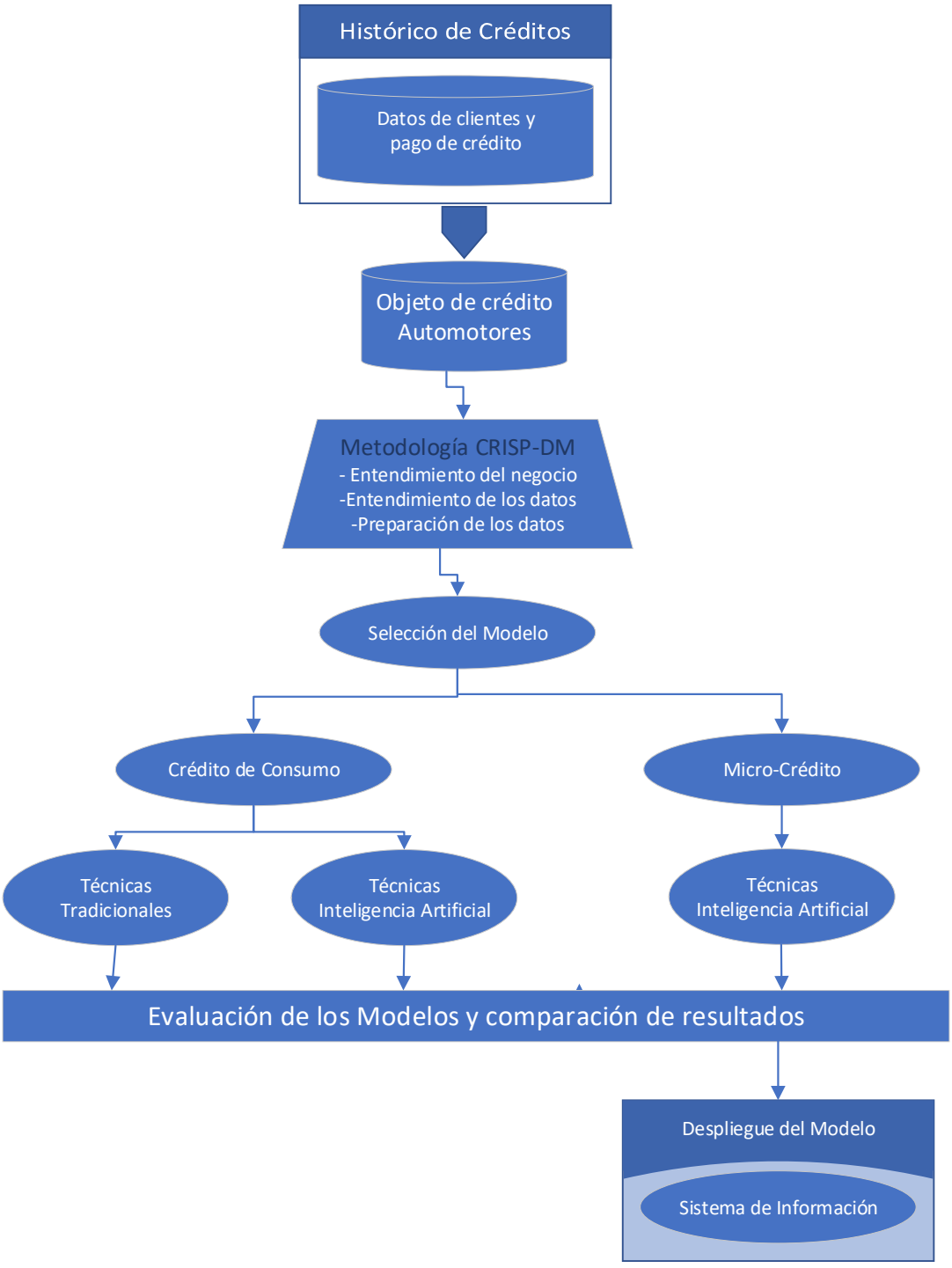
Las organizaciones mediante la aplicación de técnicas de IA podrán sumar como clientes a las personas que no están siendo atendidas por las instituciones financieras que solo usan técnicas tradicionales [57], para esto deben enmarcarse en términos de tres normas regulatorias básicas: eficiencia asignativa, equidad distributiva y autonomía del consumidor o privacidad [58].

3. Plan de Trabajo

- Diseñar un modelo de Credit Scoring que fomente la inclusión social y el uso de técnicas de inteligencia artificial explicables, haciendo énfasis en las características del bien financiado.
    - Obtener documentación sobre trabajos anteriores similares.
    - Comparar los resultados de la documentación obtenida y determinar la influencia de las diferentes técnicas en la fiabilidad del modelo.
    - Diseñar un modelo que fomente la inclusión social al sistema financiero y el uso de datos no privados.
- Entrenar un modelo de *Credit Scoring* para la determinación de la probabilidad de pago de los solicitantes de un crédito automotriz con datos obtenidos del comportamiento de pago de los clientes de un banco ecuatoriano.
    - Preparar los datos: limpieza e integración de datos y construcción de nuevas variables.
    - Diseñar y obtener las muestras de datos equilibradas.
    - Discriminar datos para entrenamiento y prueba.
    - Obtener un set de 3 modelos que combinados obtengan un solo resultado confiable y que al menos uno de ellos sea un Árbol de Decisión.
    - Analizar los resultados obtenidos.
- Crear de un Sistema de Información donde se visualicen y analicen sus resultados.
    - Diseñar el ETL y crear el data warehouse, repositorio base para el Sistema de Informacion.
    - Despliegue de los resultados del modelo y diseño de gráficos que representen esos resultados.
    - Plan de implementación.
    - Plan de monitoreo y mantenimiento.
    - Documentar el proceso, la metodología y los resultados obtenidos

Para una mejor compresión del plan de trabajo, presentamos en el Diagrama 1: Plan de Trabajo, un flujo que resume las etapas de este plan.

Diagrama 1: Plan de Trabajo



3.1 Arquitectura y herramientas utilizadas para el diseño del modelo.

La arquitectura que utilizamos para el diseño del modelo de Credit Scoring es la que se muestra en la Figura 5:

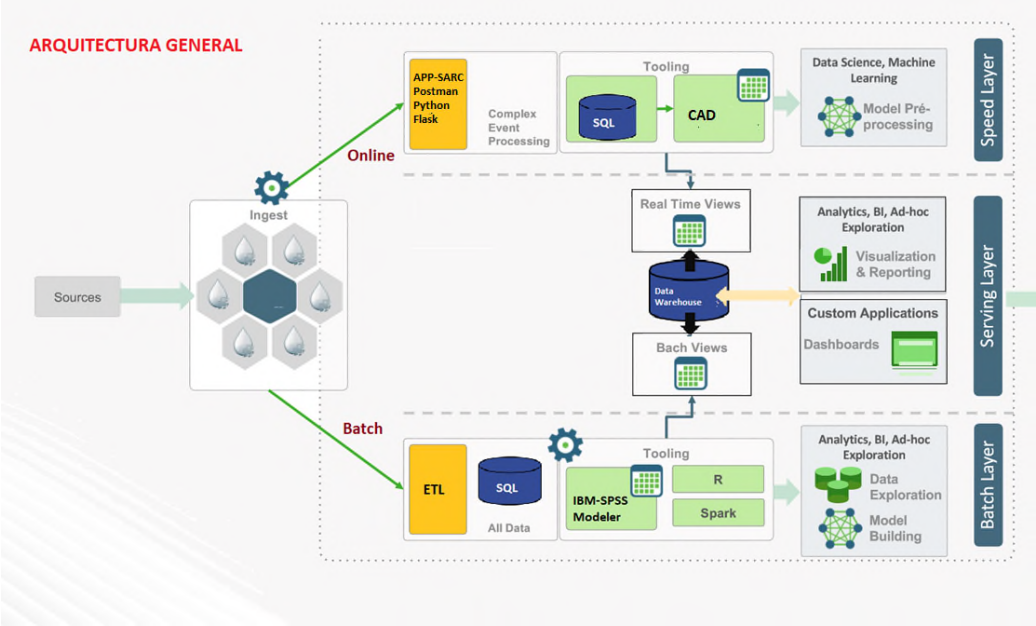


Figura 5: Arquitectura utilizada para el diseño del modelo de Credit Scoring

Como se puede observar esta arquitectura contempla dos partes:

- Arquitectura para la construcción y desarrollo de los modelos lo cual se lo hace utilizando datos históricos, por eso la palabra Batch que en español significa por lotes, por lo tanto, no es en linea.
- Arquitectura para la puesta en producción de los modelos lo cual se lo hace en línea

3.1.1 Arquitectura para construcción de los modelos (Batch)

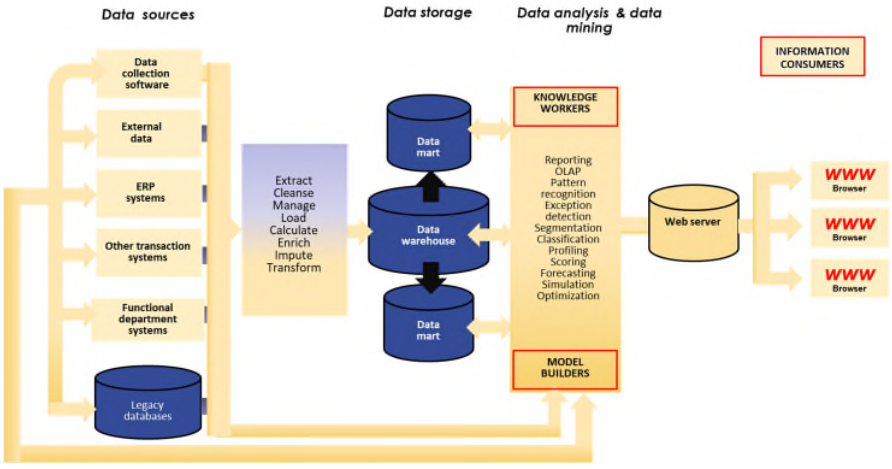


Figura 6: Arquitectura para la construcción de los modelos Batch

Con la arquitectura mostrada en la Figura 6 se realizan las etapas de: Comprensión del Negocio, Comprensión y Preparación de los Datos, Modelado, y Evaluación.

3.1.2 Arquitectura para Despliegue de los modelos (On-line)

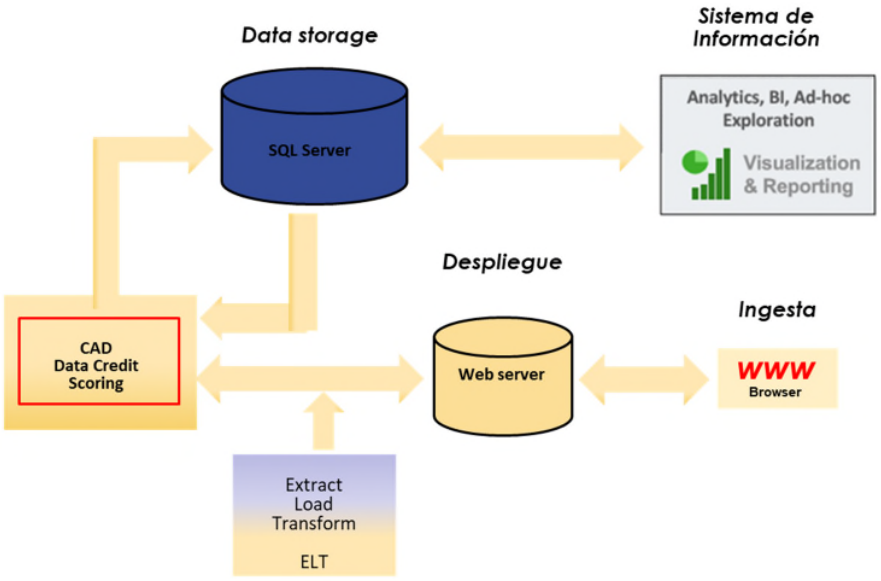


Figura 7: Arquitectura para el despliegue de los modelos On Line

Tal como se muestra en la Figura 7, para la *ingesta*, así llamada a la carga cuando el procedimiento es en línea, limpieza, procesamiento y almacenamiento de los datos de las solicitudes, vamos a utilizar la herramienta *Collaboration and Deployment (CAD),* descrita más adelante en *IBM SPSS Collaboration and Deployment* (CAD) .

Esta arquitectura por medio de un servicio web provee una calificación en línea para las solicitudes de crédito, integrándose con la aplicación del banco que maneja el flujo de estas solicitudes en los centros de venta de autos.

La aplicación informática que el banco entrega a las concesionarias permite el ingreso de información básica de los prospectos, información a partir de la cual, por medio de consultas a fuentes de información, entregan al sistema de puntuación los datos requeridos por el modelo.

Las fuentes de información son empresas dedicadas a almacenar y distribuir datos personales, se asume que este servicio cumple con la Ley de protección de datos.

4. Metodología utilizada en el diseño y despliegue

Para obtener los resultados esperados, vamos a ejecutar una metodología de Minería de Datos con IA que aseguren la integridad de los datos y den un aval científico y profesional sobre los resultados[59].

La metodología que seguiremos es Cross-Industry Standard Process for Data Mining cuyas siglas son **CRISP-DM**, que es un método probado para orientar los trabajos de minería de datos.

- Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.
- Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.

La metodología que seguimos se llama CRISP-DM y se ilustra en la Figura 8.

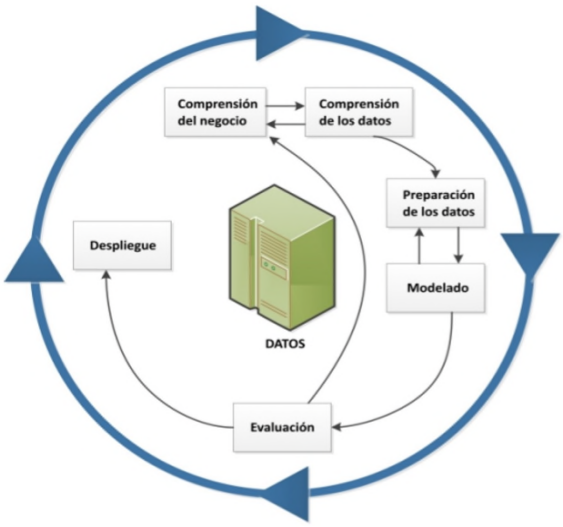


Figura 8: Ciclo de vida de minería de datos

Esta es una metodología que tiene las siguientes características:

- Sin propietario.
- Neutral en aplicaciones e Industrias.
- Puede ser desarrollada sobre cualquier herramienta de D.M.
- Se enfoca en problemas de negocio y análisis técnico.
- Proceso viable y repetible por personas con poca experiencia en D.M
- Permite plasmar las experiencias del análisis con el fin de replicarlas.

- Ayuda en la planeación y administración de un proyecto de D.M.
- Dado que es una metodología abierta permite empatar trabajos con distintos consultores.

El modelo de CRISP-DM es flexible y se puede personalizar fácilmente, permitiendo crear modelos de minería de datos que se adapten a necesidades concretas.

El ciclo vital del modelo contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases. La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario.

Sus fases se describen a continuación

Comprensión del negocio

Probablemente la fase más importante del proceso de minería de datos es la comprensión del negocio, que contiene la determinación de objetivos comerciales, la evaluación de la situación, la determinación de los objetivos de la minería de datos y la producción de un plan del proyecto.

- Evolución del Análisis del Negocio
    - Análisis Predictivo Vs OLAP
- Objetivos del negocio
    - Situación de la compañía o institución.
- Valoración actual del negocio
    - Inventario de recursos.
    - Supuestos y restricciones.
    - Relación Costo/Beneficio.
- Entender qué es Minería de Datos
    - Objetivos del proyecto de MD
        - Objetivos del proyecto.
        - Criterios de éxito para el proyecto
    - Elaborar plan de trabajo
        - Etapas del proyecto.
        - Duración por etapa.
        - Recursos requeridos.
        - Dependencias que se involucran.
        - Tener la herramienta adecuada.

Comprensión de los datos

Los datos proporcionan el "material sin procesar" de la minería de datos. La fase de comprensión de los datos está dirigida a cubrir la necesidad de comprender cuáles son los orígenes de los datos y las características de dichos orígenes. Incluye la recopilación de los datos iniciales, la descripción, exploración y verificación de la calidad de datos.

- Evolución del Análisis de los datos
- Recolección inicial de datos.
    - Listado y ubicación de la fuentes de datos.
- Descripción de datos.
    - Cantidad de campos.
    - Cantidad de registros.
    - Formato de los campos.
    - Tipo de variables.
- Análisis Exploratorio de datos
    - Reportes de exploración.
        - Estadísticos descriptivos.
        - Gráficos.
- Calidad de los datos
    - Evaluar completitud de los datos.

Preparación de los datos

Después de catalogar los orígenes de los datos, será necesario preparar los datos para su análisis. La fase de preparación de los datos incluye la selección, limpieza, construcción, integración y asignación de formato de los datos.

- Selección de datos.
    - Selección de campos y registros.
- Limpieza de datos.
    - Eliminar registros duplicados.
    - Eliminar registros inconsistentes.
- Construcción de data.
    - Derivar nuevas variables.
    - Recodificar variables.
- Integración de datos.
    - Agregar registros.
    - Agregar variables.

Modelado

La fase del modelado se trata, obviamente, de la parte más llamativa de la minería de datos, en la que se utilizan sofisticados métodos de análisis para extraer la información de los datos. Esta fase implica la selección de las técnicas de modelado, la generación de diseños de comprobación y la generación de modelos de evaluación.

- Seleccionar la técnica adecuada.
    - Modelos no supervisado.
    - Modelos supervisado.
    - Modelos cuasi-supervisados.
- Pruebas previas con el modelo.
    - Piloto.
- Construcción del modelo
- Situación del modelo.
    - Cualidades de los modelos.
    - Revisión de parámetros.

Evaluación de resultados

Una vez elegidos los modelos, ya estaríamos preparados para evaluar la forma en que los resultados del análisis pueden ayudarnos a lograr los objetivos comerciales. Los elementos principales de la fase de evaluación de resultados son la revisión del proceso de minería de datos y la determinación de los siguientes pasos.

- Evaluar los resultados
    - Respecto al objetivo del negocio
    - Respecto a los criterios de éxito del proyecto de D.M.
- Revisar el proceso completo
    - Revisar la metodología completa
- Determinar los pasos a seguir.
    - Lista de acciones o planes a desarrollar.

Despliegue

La fase despliegue se centra en la integración de nuevos conocimientos dados en el proceso comercial diario a fin de resolver el problema original comercial. Esta fase incluye el despliegue, el control y el mantenimiento del plan, la producción de un informe final, así como la revisión del proyecto.

- Plan de implementación.
- Plan de monitoreo y mantenimiento.
- Determinar los pasos a seguir.

    - Lista de acciones o planes a desarrollar.
- Reporte final.
- Revisión del proyecto de D.M

Esta metodología fue aplicada en los siguientes procedimientos:

- Recepción de los datos con el diccionario de variables.
- Preparación de los datos, así como procedimientos de validación del diccionario de datos y preprocesamiento de los datos para que, una vez los datos estén organizados, se puedan utilizar en un modelado.
- Análisis la dispersión de la información, y establecer mecanismos para concentrar en un repositorio común a todas las fuentes de datos disponibles en un mismo formato.
- Disponibilidad de la información en un mismo formato, se realizaron procesos exploratorios para que, mediante la aplicación de técnicas estadísticas, se identifiquen datos atípicos y/o registros que sean muy distintos a los demás de sus segmentos.
- Identificación de las variables más importantes que intervienen en el cumplimiento o no de los objetivos establecidos.
- Aplicación de técnicas estadísticas que encuentran relaciones, patrones de comportamiento y tendencias de los datos.
- Realización de reuniones con todas las áreas involucradas, para analizar los métodos y procedimientos que actualmente se realizan para el análisis de los clientes.
- Puesta en producción de un entorno de monitoreo, calificación y análisis de los resultados de los modelos, el mismo que tiene características funcionales que facilitarán la compresión de las gestiones que la entidad financiera realiza.

5. Diseño del modelo siguiendo la metodología CRISP-DM

Para el diseño del modelo se seguirá estrictamente la metodología CRISP-DM y, si bien el objetivo es construir un modelo de Credit Scoring con énfasis en las características del bien financiado, para que fomente la inclusión social al sistema financiero, basado en técnicas de Inteligencia Artificial explicables para que en su puesta en producción habilite la creación de un Sistema de Información donde se visualicen y analicen sus resultados. Por otro lado, también debemos demostrar que el modelo que obtuvimos es mejor que los construidos con técnicas tradicionales y que la diferencia de precisión y confianza de este modelo respecto a técnicas de IA, denominados cajas negras, no es significativa.

Para ello inicialmente construiremos un modelo desarrollado con técnicas tradicionales, luego utilizando los mismos datos, es decir aprovechando la preparación de datos realizada, construiremos algunos modelos utilizando los *Nodos de modelado automático 1: Clasificador automático*, para comparar los modelos seleccionados por esta técnica con los Árboles de Decisión y determinar si su utilización es apropiada con el fin de cumplir los objetivos de este proyecto.

5.1 Modelo para financiamiento Automotriz – Consumo

5.1.1 Entendimiento del negocio

Los ***Créditos de consumo*** son préstamos que concede una institución financiera para la adquisición de bienes o servicios. Es decir, recoge los créditos otorgados para compras comunes, como la compra de un automóvil, muebles, viajes, cualquier otro gasto extra o imprevisto.

La característica más importante es que para su otorgamiento y determinación de monto, cuotas, etc., depende de la capacidad de pago de los prospectos, la misma que está determinada por la disponibilidad de los ingresos menos los gastos y pagos con entidades financieras.

En este tipo de créditos, la contribución de un modelo en el otorgamiento es bastante limitada, por un lado, si existe disponibilidad para el pago de un crédito seguramente el modelo aprobará la operación sin importar el destino del crédito, y por otro lado si el prospecto no tiene capacidad de pago o no puede justificar el compromiso de pago, aunque haya tenido un buen récord crediticio, el modelo no le otorgará el crédito. Estas situaciones no generan inclusión financiera, deficiencia que se pretende mejorar con el desarrollo del modelo propuesto.

La definición de este tipo de créditos, en el Ecuador, está dada en el glosario de términos, en la parte correspondiente al Crédito en ***Créditos de consumo***.

5.1.2 Entendimiento de los datos

Para comprender la estructura, generación y manipulación de los datos involucrados, fue necesario hacer un inventario de éstos, en el orden que se describe a continuación en la Tabla 6.

Inventario de datos

Tabla 6: Inventario de datos

No.	Nombre archivo	Descripción
1	Clientes	Información detallada en la ficha del cliente (Personas, Empresas)
2	Crédito	Información detallada de las operaciones de crédito (Personas, Empresas)
3	Dato maestro de artículos	Información relacionada a la clasificación de los artículos
4	Cartera	Información de cartera por cliente
5	Condición de pago	Codificación de condiciones de pago
6	Solicitudes de crédito	Información de solicitudes pendientes, rechazadas, etc.

Variables relacionadas con el cliente

Nombre de la variable	Nombre en el modelo	Tipo	Descripción
Sexo del cliente	codigoSexoCliente	VARCHAR	M, F
Edad del Cliente	edadCliente	ENTERO	Rango [23 – 93]
Estado Civil del cliente	codigoEstadoCivilCliente	VARCHAR	C, D, S, U, V
Nivel de estudios del cliente	codigoNivelEstudiosCliente	VARCHAR	G, N, P, S, U
Provincia de residencia del Cliente	codigoProvinciaCliente	VARCHAR	01,02,03,04,05,06,07,08,09,10,12, 13,14,15,16,17,18,21,22,23,24
Ciudad de residencia del Cliente	codigoCiudadCliente	VARCHAR	01.08, 02.03, 07.05, 08.07, 09.03, 09.11, 11.100, 12.070, 14.030, 14.120, 19.040, 21.010, 01.04, 03.03, 04.01, …
Tipo de residencia del cliente	tipoResidenciaCliente	VARCHAR	A, F, N, P, S
Ingresos del cliente	ingresosCliente	DECIMAL	Rango [0 – 40000]
Gastos del cliente	gastosCliente	DECIMAL	Rango [0 – 7410.45]

Variables relacionadas al crédito u operación

Nombre de la variable	Nombre en el modelo	Tipo	Descripción
Tipo de Crédito	tipoOperacion	VARCHAR	LP, LT, PT
Porcentaje de la Entrada	porcentajeEntradaOperacion	DECIMAL	Rango [0.0 % – 0.734 %]
Monto del Préstamo	montoOperacion	DECIMAL	Rango [3748.4 – 39956.10]
Número de cuotas	numeroCuotasOperacion	ENTERO	Rango [1 – 120]
Monto de la cuota	montoCuotaOperacion	DECIMAL	Rango [15.715 – 4018.095]
Plazo del préstamo	plazoOperacion	ENTERO	Rango [1 – 60]
Tasa préstamo	tasaOperacion	DECIMAL	Rango [11.23 % – 21.70 %]

Variable: Tipo de Crédito

- LP Credito de vehículos livianos para uso particular
- LT Credito de vehículos livianos para uso de trabajo
- PT Credito de vehículos pesados para uso de trabajo
-

Variables relacionadas con el vehículo

Variable: Marca del Vehículo, Clase de Vehículo y Estado del vehículo

Nombre de la variable	Nombre en el modelo	Tipo de variable	Descripción
Valor del vehículo	valorVehiculo	ENTERO	Rango [6500 – 60000]
Marca del Vehículo	marcaVehiculo	VARCHAR	RENAULT, HYUNDAI, NISSAN, KIA, CHEVROLET, MAZDA, TOYOTA, HONDA, SUZUKI, WOLKSWAGEN, MITSUBISHI, JEEP, PEUGEOT, HINO, FORD, GREATWALL, BYD, FIAT, DFSK, CHERY, CHANGHE, OTROS.
Clase del Vehículo	claseVehiculo	VARCHAR	AUTOMOVIL, BUS, CAMION, CAMIONETA, FURGONETA, JEEP, TRACTOR, OTROS.
Estado del vehículo	estadoVehiculo	VARCHAR	NUEVO, USADO

5.1.3 Preparación de los datos

Para la construcción del score de crédito fue necesario diseñar una ***Base de datos*** que almacene las ***Variables Cualitativas*** y ***Variables Cuantitativas***..

Análisis preliminar de los datos

Para el diseñó de la base de datos se hizo un análisis preliminar que constó de cuatro pasos:

a) Verificar la integridad de los datos, es decir que correspondan a datos reales, verdaderos y que contengan toda la información de las solicitudes de crédito que fueron recibidas.
b) Transformar variables y tratar los datos nulos. En el primer caso se construyeron nuevas variables a través de cálculos, combinaciones o cruces. En el segundo caso se analizó el número de datos perdidos y el grado de relevancia de la variable para buscar una regla de asignación o simplemente eliminarla.
c) Segmentar los casos de cada variable; consistió en agrupar por categorías. Por ejemplo, segmentar por provincia, género, nivel de ingresos, rangos de edades, entre otros.
d) Seleccionar preliminarmente las variables; esta actividad se realizó en función del nivel de correlación que tenga una variable con otra, o evaluando el nivel de predicción que tenga la variable sobre el caso de estudio.

Selección de la muestra

El desarrollo de la puntuación de crédito requiere de una muestra extensa que contenga información relevante de clientes y créditos, de tal manera que se pueda identificar factores que influyan en la probabilidad de incumplimiento. Preparamos la muestra para que posea las siguientes características:

- Ser representativa de clientes potenciales.
- Poseer información suficiente sobre el comportamiento de pago.
- Contener información de buenos y malos clientes.
- Determinar el rango de tiempo en el que se desea analizar el desempeño del cliente.

La muestra que se ha recogido consiste en aproximadamente 24.920 operaciones, las cuales contienen toda la información correspondiente en las siguientes tablas:

- Clientes,
- Cartera desde 2010,
- Detalle de las cuotas de cada operación, y
- Características del automóvil.

En la Figura 9, se muestra la ruta creada para obtener la muestra

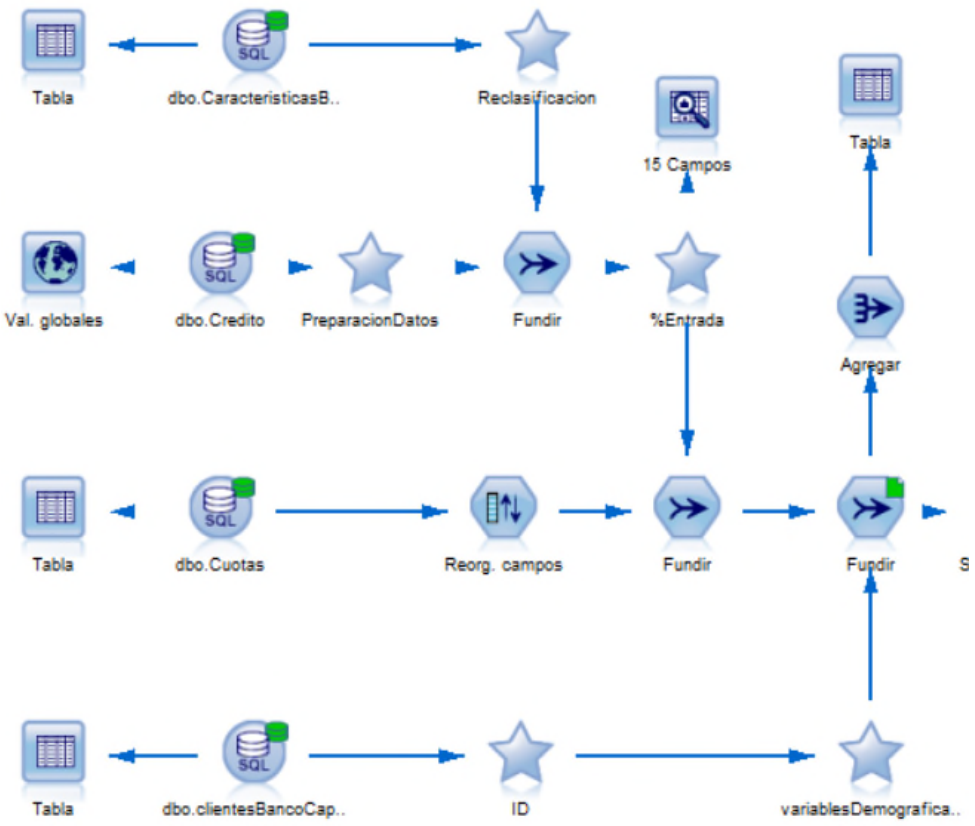


Figura 9: Ruta para obtener la muestra

En la Figura 10, se muestra el resultado de los registros seleccionas de la muestra en una tabla con 24.920 registros.

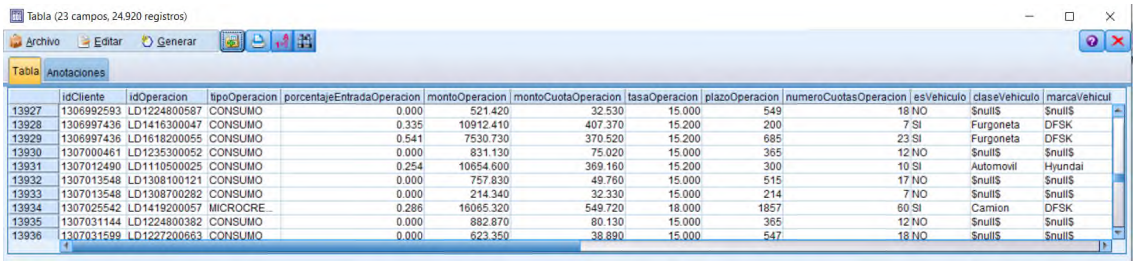


Tabla (23 campos, 24.920 registros)

Archivo Editar Generar

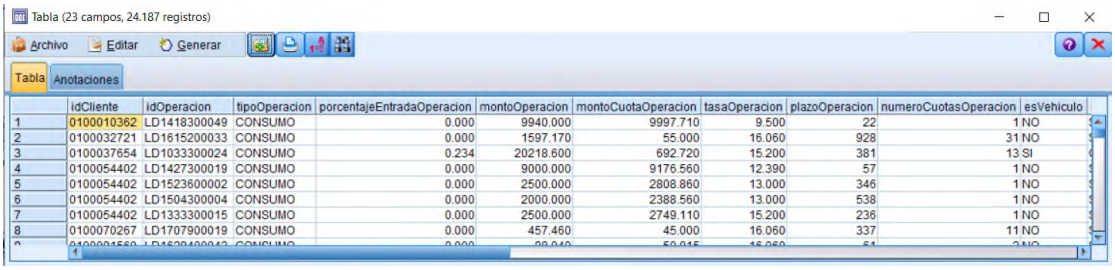
Tabla Anotaciones

	idCliente	idOperacion	tipoOperacion	porcentajeEntradaOperacion	montoOperacion	montoCuotaOperacion	tasaOperacion	plazoOperacion	numeroCuotasOperacion	esVehiculo	claseVehiculo	marcaVehicul
13927	1306992593	LD1224800587	CONSUMO	0.000	521.420	32.530	15.000	549	18	NO	$null$	$null$
13928	1306997436	LD1416300047	CONSUMO	0.335	10912.410	407.370	15.200	200	7	SI	Furgoneta	DFSK
13929	1306997436	LD1618200055	CONSUMO	0.541	7530.730	370.520	15.200	685	23	SI	Furgoneta	DFSK
13930	1307000461	LD1235300052	CONSUMO	0.000	831.130	75.020	15.000	365	12	NO	$null$	$null$
13931	1307012490	LD1110500025	CONSUMO	0.254	10654.600	369.160	15.200	300	10	SI	Automovil	Hyundai
13932	1307013548	LD1308100121	CONSUMO	0.000	757.830	49.760	15.000	515	17	NO	$null$	$null$
13933	1307013548	LD1308700282	CONSUMO	0.000	214.340	32.330	15.000	214	7	NO	$null$	$null$
13934	1307025542	LD1419200057	MICROCRE..	0.286	16065.320	549.720	18.000	1857	60	SI	Camion	DFSK
13935	1307031144	LD1224800382	CONSUMO	0.000	882.870	80.130	15.000	365	12	NO	$null$	$null$
13936	1307031599	LD1227200663	CONSUMO	0.000	623.350	38.890	15.000	547	18	NO	$null$	$null$

Figura 10: Registros seleccionados (muestra)

Posteriormente, se descartaron los créditos que fueron parte de una cartera vendida sin que se haya terminado el plazo para su cobro, ya que estos no tuvieron una culminación registrada en el banco que permita establecer su real comportamiento.

Después de ello quedaron 24.187 créditos tal como lo muestra la Figura 11.



	idCliente	idOperacion	tipoOperacion	porcentajeEntradaOperacion	montoOperacion	montoCuotaOperacion	tasaOperacion	plazoOperacion	numeroCuotasOperacion	esVehiculo
1	0100010362	LD1418300049	CONSUMO	0.000	9940.000	9997.710	9.500	22	1	NO
2	0100032721	LD1615200033	CONSUMO	0.000	1597.170	55.000	16.060	928	31	NO
3	0100037654	LD1033300024	CONSUMO	0.234	20218.600	692.720	15.200	381	13	SI
4	0100054402	LD1427300019	CONSUMO	0.000	9000.000	9176.560	12.390	57	1	NO
5	0100054402	LD1523600002	CONSUMO	0.000	2500.000	2808.860	13.000	346	1	NO
6	0100054402	LD1504300004	CONSUMO	0.000	2000.000	2388.560	13.000	538	1	NO
7	0100054402	LD1333300015	CONSUMO	0.000	2500.000	2749.110	15.200	236	1	NO
8	0100070267	LD1707900019	CONSUMO	0.000	457.460	45.000	16.060	337	11	NO

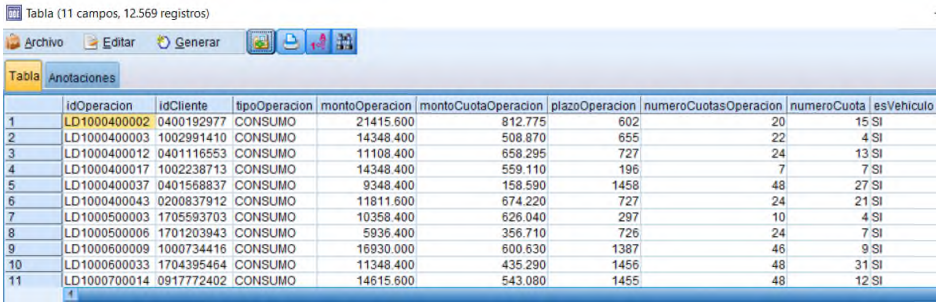
Figura 11: Registros con comportamiento de pago

Finalmente se descartaron también aquellos créditos que, a la fecha de construcción del modelo (febrero 2019), tenían cuotas por vencer, es decir las operaciones que a esa fecha estaban culminadas, por lo que tampoco se podría saber su comportamiento final, quedan al final 22.407 créditos, tal como lo muestra la Figura 12.

	idCliente	idOperacion	tipoOperacion	montoOperacion	montoCuotaOperacion	tasaOperacion	plazoOperacion	porcentajeEntradaOperacion	numeroCuotasOperacion	claseVehiculo	marcaVehiculo
1	0400192977	LD1000400002	CONSUMO	21415.600	812.775	17.000	602	0.286	20	Camioneta	Mazda
2	1002991410	LD1000400003	CONSUMO	14348.400	508.870	17.000	655	0.282	22	Jeep	Hyundai
3	0401116553	LD1000400012	CONSUMO	11108.400	658.295	17.000	727	0.337	24	Automovil	Renault
4	1002238713	LD1000400017	CONSUMO	14348.400	559.110	17.000	196	0.244	7	Camioneta	Mazda
5	1702151471	LD1000400027	CONSUMO	4018.200	253.325	15.200	546	0.000	18	$null$	$null$
6	0401568837	LD1000400037	CONSUMO	9348.400	158.590	17.000	1458	0.220	48	Automovil	Renault
7	0200837912	LD1000400043	CONSUMO	11811.600	674.220	17.000	727	0.421	24	Camioneta	Chevrolet
8	1716860695	LD1000400045	CONSUMO	4000.000	251.920	15.200	546	0.000	18	$null$	$null$

Figura 12: Registros con operaciones culminadas a febrero 2019

Como se trata de un modelo de score de crédito automotriz, seleccionamos solo aquellos créditos pertenecientes a esta categoría, finalmente quedan 12.569 créditos, tal como lo muestra la Figura 13.

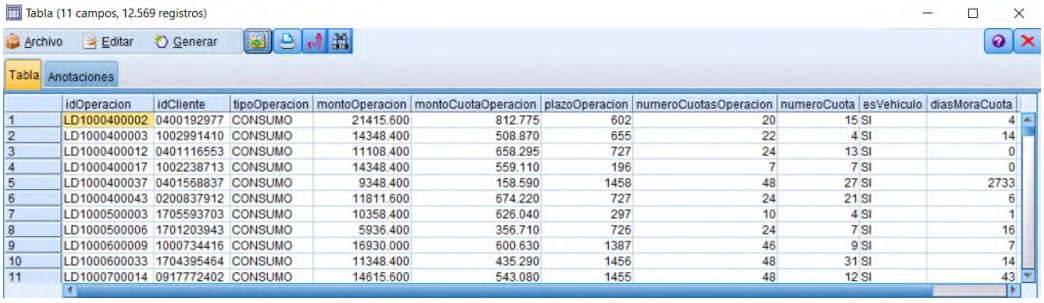


	idOperacion	idCliente	tipoOperacion	montoOperacion	montoCuotaOperacion	plazoOperacion	numeroCuotasOperacion	numeroCuota	esVehiculo
1	LD1000400002	0400192977	CONSUMO	21415.600	812.775	602	20	15	SI
2	LD1000400003	1002991410	CONSUMO	14348.400	508.870	655	22	4	SI
3	LD1000400012	0401116553	CONSUMO	11108.400	658.295	727	24	13	SI
4	LD1000400017	1002238713	CONSUMO	14348.400	559.110	196	7	7	SI
5	LD1000400037	0401568837	CONSUMO	9348.400	158.590	1458	48	27	SI
6	LD1000400043	0200837912	CONSUMO	11811.600	674.220	727	24	21	SI
7	LD1000500003	1705593703	CONSUMO	10358.400	626.040	297	10	4	SI
8	LD1000500006	1701203943	CONSUMO	5936.400	356.710	726	24	7	SI
9	LD1000600009	1000734416	CONSUMO	16930.000	600.630	1387	46	9	SI
10	LD1000600033	1704395464	CONSUMO	11348.400	435.290	1456	48	31	SI
11	LD1000700014	0917772402	CONSUMO	14615.600	543.080	1455	48	12	SI

Figura 13: Registros de crédito automotriz

Creación de variables adicionales

Una de las variables más importantes, que deben ser creadas, es diasMoraCuota, tal como lo muestra la Figura 14 en la última columna, que nos indica los días de mora que tuvo cada cuota desde el vencimiento hasta día de pago o en el caso de que esté vencida pero aún no registra el pago total, la mora corresponde a los días vencidos a la fecha del análisis.



	idOperacion	idCliente	tipoOperacion	montoOperacion	montoCuotaOperacion	plazoOperacion	numeroCuotasOperacion	numeroCuota	esVehiculo	diasMoraCuota
1	LD1000400002	0400192977	CONSUMO	21415.600	812.775	602	20	15	SI	4
2	LD1000400003	1002991410	CONSUMO	14348.400	508.870	655	22	4	SI	14
3	LD1000400012	0401116553	CONSUMO	11108.400	658.295	727	24	13	SI	0
4	LD1000400017	1002238713	CONSUMO	14348.400	559.110	196	7	7	SI	0
5	LD1000400037	0401568837	CONSUMO	9348.400	158.590	1458	48	27	SI	2733
6	LD1000400043	0200837912	CONSUMO	11811.600	674.220	727	24	21	SI	6
7	LD1000500003	1705593703	CONSUMO	10358.400	626.040	297	10	4	SI	1
8	LD1000500006	1701203943	CONSUMO	5936.400	356.710	726	24	7	SI	16
9	LD1000600009	1000734416	CONSUMO	16930.000	600.630	1387	46	9	SI	7
10	LD1000600033	1704395464	CONSUMO	11348.400	435.290	1456	48	31	SI	14
11	LD1000700014	0917772402	CONSUMO	14615.600	543.080	1455	48	12	SI	43

Figura 14: Creación de la variable días de mora por cuota

La variable carteraVencidaSinCuotasPorVencer, que se describe en la Figura 15, indica si un crédito que ya debió haber sido pagado aun tiene cuotas vencidas, la creación de esta variable es muy importante para determinar qué operaciones fueron buenas o malas.

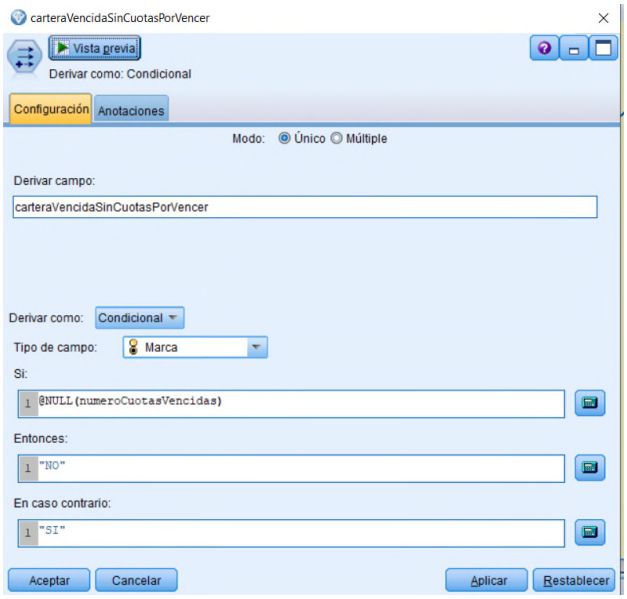


Figura 15: Definición de variable: carteraVencidaSinCuotasPorVencer

Definición de clientes buenos y malos

Para clasificar a un cliente como bueno o malo analizamos su comportamiento de pago: número de cuotas vencidas, monto por pagar y número de veces que cayó en mora; además, tomamos en cuenta el tipo de crédito y los factores que pueden influir en el comportamiento de pago, así como las políticas y objetivos de reducción de riesgo adoptadas por cada institución financiera.

De acuerdo con el tipo de cartera se considerará, en mora a los quince o treinta días de vencido el pago de la cuota. Mientras mayor o menor sean los días para considerar en mora, el rango de aceptación o rechazo subirá o bajará. Por ejemplo: cuando se establece 15 días para considerar la cartera en mora, la tasa de rechazo se incrementará; por el contrario, cuando se estima vencida a los 30 días, la tasa de aprobación se incrementa.

De acuerdo con la información recibida por el banco se crea una nueva variable objetivo o dependiente que defina a las operaciones como buenas o malas.

Aquellas operaciones que fueron vendidas estando vencidas y también las castigadas, fueron declaradas como malas, adicionalmente, a esta característica se suman las operaciones que cumplen con la premisa de ser carteraVencidaSinCuotasPorVencer explicada anteriomente, estas condiciones se describe en la Figura 16 .

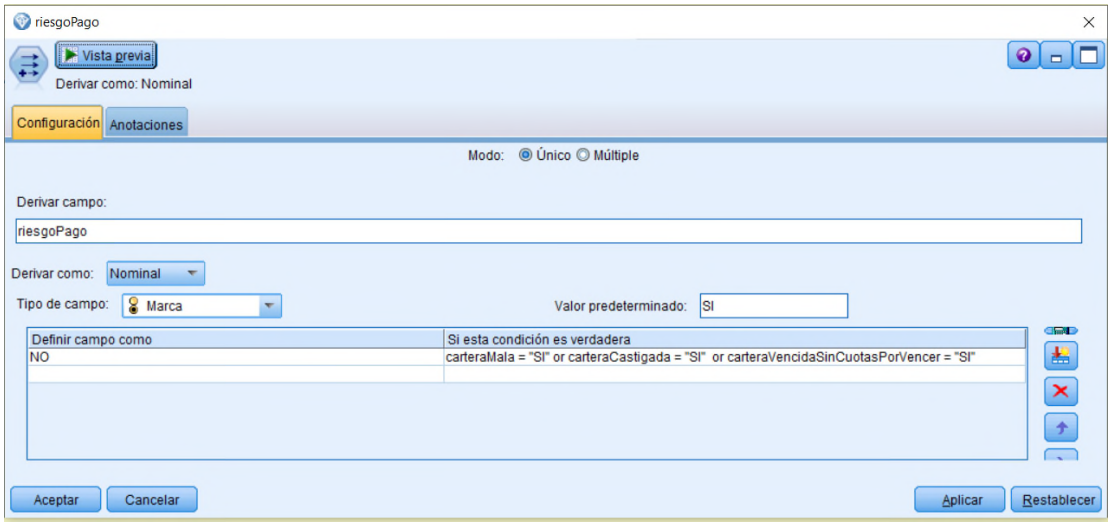


Figura 16: Definición de variable objetivo riesgo de pago

Al graficar cada una de las categorías de la variable objetivo el porcentaje de créditos buenos y malos es el que se muestra en la Figura 17.

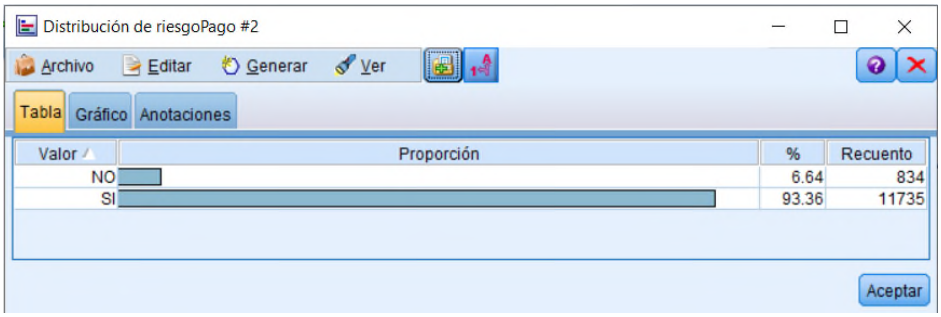


Figura 17: Distribución de las categorías de riesgo de pago

Equilibrado de la muestra

Para que el modelo aprenda sin sesgo, tanto de los buenos como de los malos créditos, se debe equilibrar el número de casos para cada uno de los valores de la clase, lo que significa que debemos tomar una muestra con números compatibles de créditos buenos y créditos malos, la ruta generada para esto es la mostrada en la Figura 18.

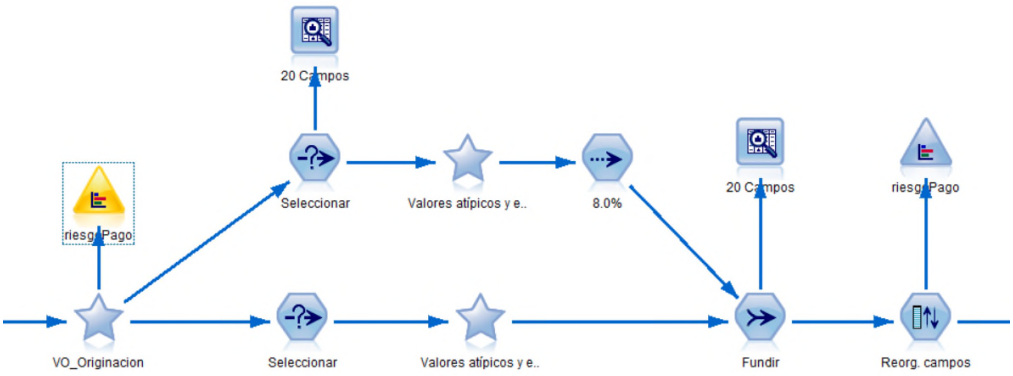


Figura 18: Ruta para equilibrar la variable objetivo

Previo a esto, de manera individual, en cada categoría de la variable dependiente se eliminaron los registros que presentaban datos faltantes o atípicos. Finalmente, se obtuvo la siguiente proporción de datos de cartera considerada buena y mala tal como se muestra en la Figura 19 en la columna "%".

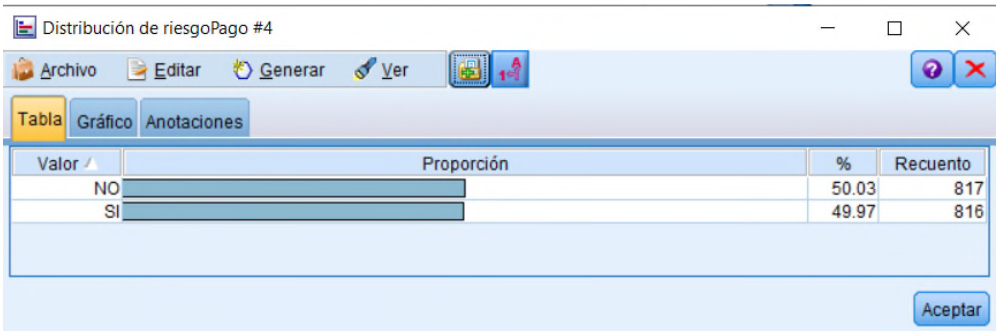


Figura 19: Distribución de las categorías de riesgo de pago

Se observa que la muestra quedo equilibrada y se puede proceder con el entrenamiento el modelo.

Partición de la muestra

En minería de datos, la clasificación se realiza en dos pasos, en primer lugar, se entrena el modelo y luego se prueba su asertividad; para lo cual los datos se dividen en dos conjuntos: el de entrenamiento y el de prueba.

Los nodos Partición de la herramienta Modeler se utilizan para generar un campo de partición que divide los datos en subconjuntos o muestras independientes para las fases de entrenamiento, comprobación y validación en la generación del modelo. Se usa una muestra para generar el modelo (Entrenamiento) y otra muestra distinta para probarlo (Prueba).

En la ruta generada para para la partición de la muestra, Figura 20, el nodo Partición es el primer nodo.

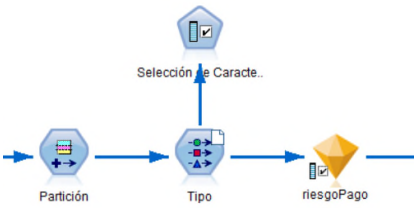


Figura 20: Nodo Partición

Las opciones de edición del nodo Partición, presentado en la Figura 21, permiten elegir el tamaño de los conjuntos de entrenamiento y prueba. Normalmente se toma aleatoriamente el 70% de los datos para entrenar, y el 30% de los datos restantes para comprobar los aciertos del modelo en la tabla de clasificación de buenos y malos en la etapa de evaluación de la metodología.

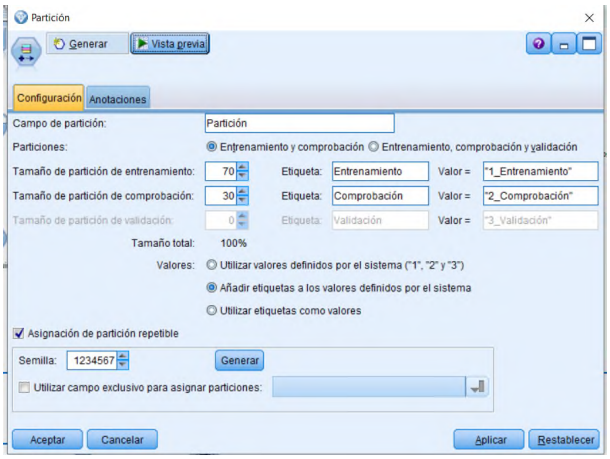


Figura 21: Configuración del nodo Partición

5.1.4 Modelado

El análisis multivariado consiste en examinar cada variable de la base de datos, de tal manera que se identifique el grado de predicción o influencia sobre la variable dependiente. En función del examen se eligen las variables que contribuirán en el desarrollo del modelo para posteriormente descubrir el algoritmo de scoring que mejor represente al caso de estudio.

Selección de variables

La Figura 22: Nodo Selección de Características, que corresponde a la continuación de la ruta de la sección anterior, se puede observar el nodo Tipo que está después del Nodo Partición.

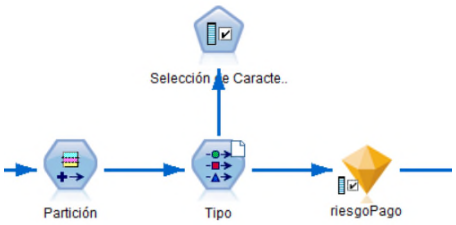
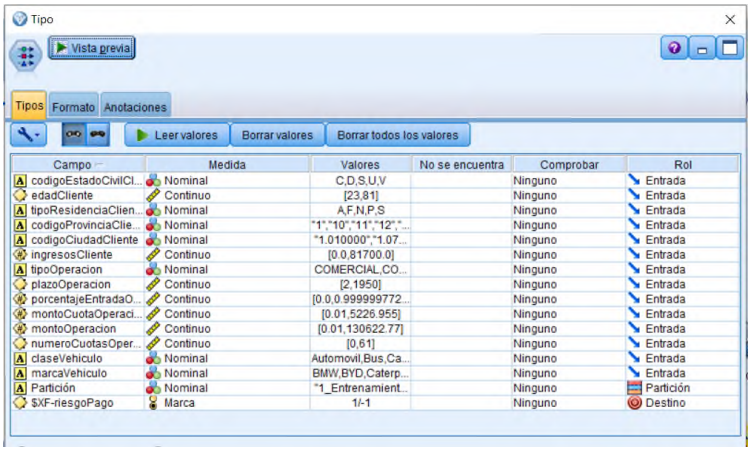


Figura 22: Nodo Selección de Características

En nodo Tipo en la Figura 23, muestra las variables que podrían participar en el modelo.

En este nodo también muestra el rol de cada variable, donde se estableció que la variable *IdOperacion* no tiene ningún rol mientras que la variable *riesgoPago* tiene el rol de *Destino*, así también para la variable *Partición* cuyo rol será *Partición*, que intervendrá como agrupadora de los datos en la evaluación del modelo.



Campo	Medida	Valores	No se encuentra	Comprobar	Rol
codigoEstadoCivilCl...	Nominal	C,D,S,U,V		Ninguno	Entrada
edadCliente	Continuo	[23,81]		Ninguno	Entrada
tipoResidenciaClien...	Nominal	A,F,N,P,S		Ninguno	Entrada
codigoProvinciaClie...	Nominal	"1","10","11","12","...		Ninguno	Entrada
codigoCiudadCliente	Nominal	"1.010000","1.07...		Ninguno	Entrada
ingresosCliente	Continuo	[0.0,81700.0]		Ninguno	Entrada
tipoOperacion	Nominal	COMERCIAL,CO...		Ninguno	Entrada
plazoOperacion	Continuo	[2,1950]		Ninguno	Entrada
porcentajeEntradaO...	Continuo	[0.0,0.999999772...		Ninguno	Entrada
montoCuotaOperaci...	Continuo	[0.01,5226.955]		Ninguno	Entrada
montoOperacion	Continuo	[0.01,130622.77]		Ninguno	Entrada
numeroCuotasOper...	Continuo	[0,61]		Ninguno	Entrada
claseVehiculo	Nominal	Automovil,Bus,Ca...		Ninguno	Entrada
marcaVehiculo	Nominal	BMW,BYD,Caterp...		Ninguno	Entrada
Partición	Nominal	"1_Entrenamient...		Ninguno	Partición
$XF-riesgoPago	Marca	1/-1		Ninguno	Destino

Figura 23: Lista de variables que participarían en el modelo

A continuación, en la ruta se editó el nodo *Selección de Características* para determinar estadísticamente el nivel de aporte que tiene cada variable para una buena discriminación de los registros en las categorías de la variable objetivo.

En el nodo *Selección de Características* en la Figura 24 se muestra cómo se configuraron las diferentes opciones para determinar estadísticamente el nivel de aporte que tiene cada variable para una buena discriminación de los registros en las categorías de la variable objetivo

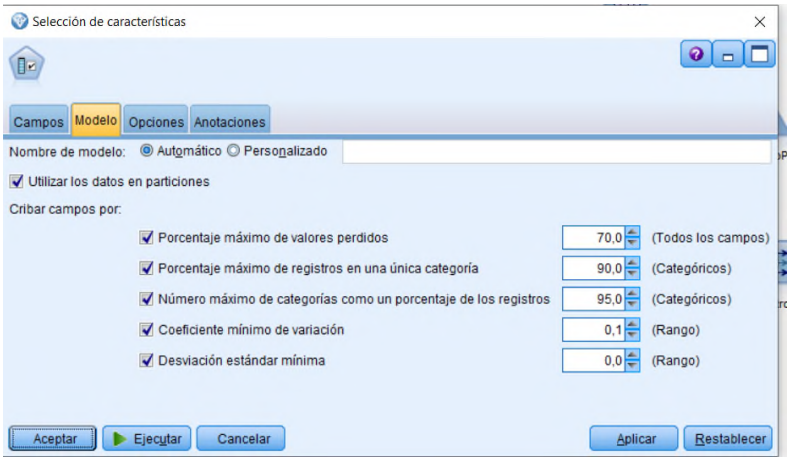


Figura 24: Configuración de nodo Selección de características

A continuación, en el Nugget de modelo *Riesgo de Pago tal como se* muestra en la Figura 25: Resultado del orden de importancia de variables, se determinó el grado de importancia que tiene cada variable para la clasificación de la variable objetivo.

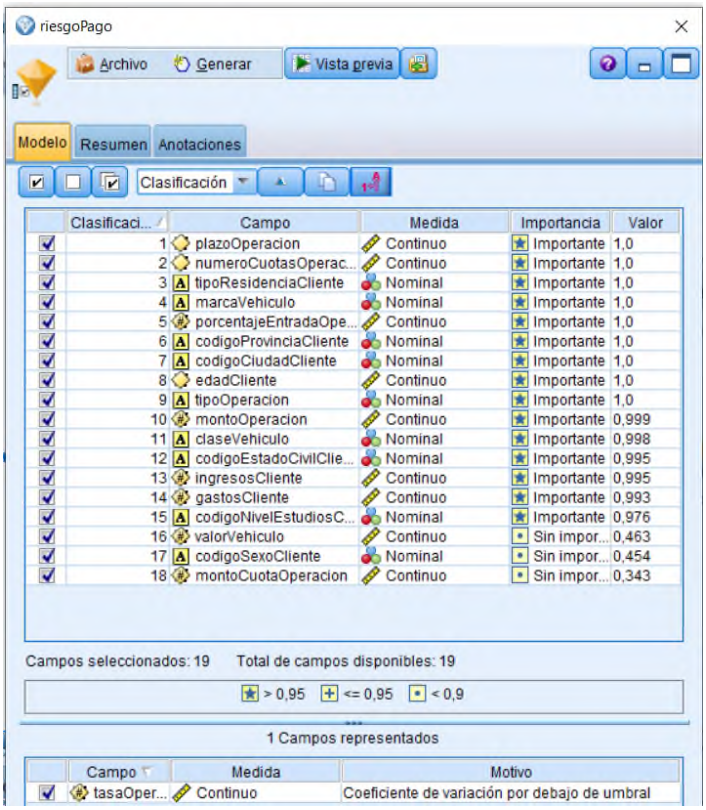


Figura 25: Resultado del orden de importancia de variables

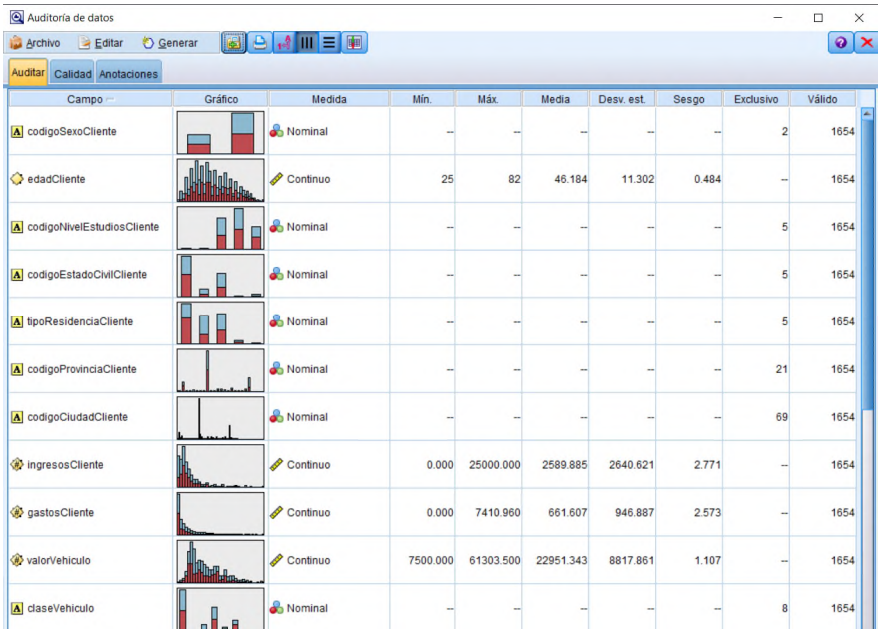
En la visualización de este nodo se observó que las 15 primeras variables son importantes mientras que, por el contrario, las demás aportan muy poco al modelo; también se concluye que la variable *tasa de operación* no da ningún aporte ya que el coeficiente de variación está por debajo del umbral. Sin embargo, por temas de integración con el proceso de calificación de crédito, hemos incorporado estas 4 variables restantes que, en

principio, no serán tomadas en cuenta por el modelo, pero como se trata de crear un modelo de inteligencia artificial, es muy probable que cuando se lo reentrene, estas variables tengan una variación suficiente para aportar al modelo, junto con nuevas variables que irán apareciendo en el camino, conforme el proceso de calificación acumule información.

En la Figura 26, se observa que a continuación del nodo *Riesgo de Pago*, se ha colocado el nodo *Auditoría de datos*, para observar la distribución de cada una de las variables con respecto a la variable *Destino*, a fin de verificar la calidad de los datos.

Figura 26: Configuración de nodo Auditoría de Datos

En el nodo *Auditoría de Datos* que se muestran en la Figura 27: Auditoria de las variables, parte 1, y Figura 28: Auditoria de las variables, parte 2, en la pestaña *Auditar*, se observa la distribución de cada una de las variables que participarán en el modelo y su repercusión sobre la variable objetivo, mostrando esta relación con diferentes colores: los créditos buenos con color rojo y los créditos malos con color azul.



Campo	Gráfico	Medida	Mín.	Máx.	Media	Desv. est.	Sesgo	Exclusivo	Válido
codigoSexoCliente		Nominal	--	--	--	--	--	2	1654
edadCliente		Continuo	25	82	46.184	11.302	0.484	--	1654
codigoNivelEstudiosCliente		Nominal	--	--	--	--	--	5	1654
codigoEstadoCivilCliente		Nominal	--	--	--	--	--	5	1654
tipoResidenciaCliente		Nominal	--	--	--	--	--	5	1654
codigoProvinciaCliente		Nominal	--	--	--	--	--	21	1654
codigoCiudadCliente		Nominal	--	--	--	--	--	69	1654
ingresosCliente		Continuo	0.000	25000.000	2589.885	2640.621	2.771	--	1654
gastosCliente		Continuo	0.000	7410.960	661.607	946.887	2.573	--	1654
valorVehiculo		Continuo	7500.000	61303.500	22951.343	8817.861	1.107	--	1654
claseVehiculo		Nominal	--	--	--	--	--	8	1654

Figura 27: Auditoria de las variables, parte 1

Campo	Gráfico	Medida	Mín.	Máx.	Media	Desv. est.	Sesgo	Exclusivo	Válido
marcaVehiculo		Nominal	--	--	--	--	--	20	1654
tipoOperacion		Nominal	--	--	--	--	--	3	1654
montoOperacion		Continuo	3848.400	40333.340	15452.965	6071.230	1.036	--	1654
tasaOperacion		Continuo	11.230	21.705	15.825	1.414	1.933	--	1654
plazoOperacion		Continuo	49	1942	1291.998	512.739	-0.699	--	1654
porcentajeEntradaOperacion		Continuo	0.000	0.733	0.319	0.111	1.249	--	1654
montoCuotaOperacion		Continuo	15.715	4110.470	570.613	306.471	3.170	--	1654
numeroCuotasOperacion		Continuo	2	63	42.108	16.714	-0.693	--	1654
riesgoPago		Marca	-1	1	--	--	--	2	1654
Partición		Nominal	--	--	--	--	--	2	1654

Figura 28: Auditoria de las variables, parte 2

Cada uno de los diagramas de las Figuras 28 y 29 fueron analizados de manera individual y a manera de ejemplo a continuación en la Figura 29 se muestra el resultado para *ingresosCliente*.

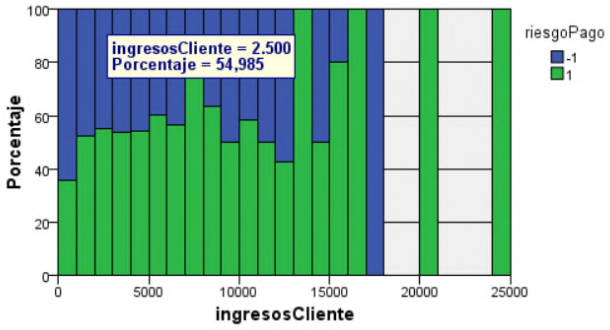
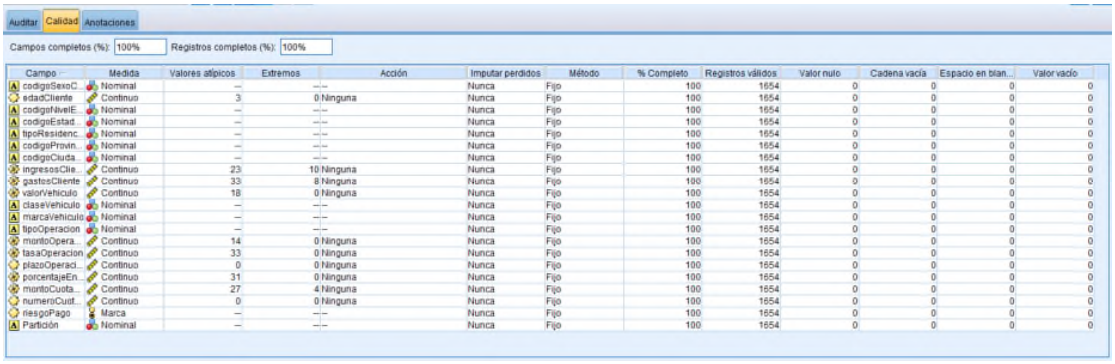


Figura 29: Resultado, variable ingresos del cliente - Normalizada.

En la Figura 29 el valor -1 corresponde a créditos malos y 1 a créditos buenos. Si la variable no aportara, todas las barras tendrían igual altura para cada uno de los colores o categorías y para esta variable no es el caso.

A continuación, en el mismo nodo *Auditoría de datos*, en la pestaña *Calidad de datos* tal como se muestra en la Figura 30, para cada variable se presentan tres aspectos: valores atípicos, valores extremos y registros completos. Los valores atípicos y extremos que se observan corresponden a registros de créditos malos que no se descartaron previamente debido a que le enseñan al modelo estas condiciones atípicas y extremas que generaron malas experiencias. En las cuatro últimas columnas de la matriz se verificó si el conjunto de datos está completo, lo cual ocurre cuando no hay valores nulos, vacíos o en blanco.



Campo	Medida	Valores atípicos	Extremos	Acción	Imputar perdidos	Método	% Completo	Registros válidos	Valor nulo	Cadena vacía	Espacio en blan...	Valor vacío
codigoSexoC...	Nominal	--	--	--	Nunca	Fijo	100	1654	0	0	0	0
edadCliente	Continuo	3	0	Ninguna	Nunca	Fijo	100	1654	0	0	0	0
codigoNivelE...	Nominal	--	--	--	Nunca	Fijo	100	1654	0	0	0	0
codigoEstad...	Nominal	--	--	--	Nunca	Fijo	100	1654	0	0	0	0
tipoResidenc...	Nominal	--	--	--	Nunca	Fijo	100	1654	0	0	0	0
codigoProvin...	Nominal	--	--	--	Nunca	Fijo	100	1654	0	0	0	0
codigoCiuda...	Nominal	--	--	--	Nunca	Fijo	100	1654	0	0	0	0
ingresosClie...	Continuo	23	10	Ninguna	Nunca	Fijo	100	1654	0	0	0	0
gastosCliente	Continuo	33	8	Ninguna	Nunca	Fijo	100	1654	0	0	0	0
valorVehiculo	Continuo	18	0	Ninguna	Nunca	Fijo	100	1654	0	0	0	0
claseVehiculo	Nominal	--	--	--	Nunca	Fijo	100	1654	0	0	0	0
marcaVehiculo	Nominal	--	--	--	Nunca	Fijo	100	1654	0	0	0	0
tipoOperacion	Nominal	--	--	--	Nunca	Fijo	100	1654	0	0	0	0
montoOpera...	Continuo	14	0	Ninguna	Nunca	Fijo	100	1654	0	0	0	0
tasaOperacion	Continuo	33	0	Ninguna	Nunca	Fijo	100	1654	0	0	0	0
plazoOperaci...	Continuo	0	0	Ninguna	Nunca	Fijo	100	1654	0	0	0	0
porcentajeEn...	Continuo	31	0	Ninguna	Nunca	Fijo	100	1654	0	0	0	0
montoCuota...	Continuo	27	4	Ninguna	Nunca	Fijo	100	1654	0	0	0	0
numeroCuot...	Continuo	0	0	Ninguna	Nunca	Fijo	100	1654	0	0	0	0
riesgoPago	Marca	--	--	--	Nunca	Fijo	100	1654	0	0	0	0
Partición	Nominal	--	--	--	Nunca	Fijo	100	1654	0	0	0	0

Figura 30: Auditoría de datos: Calidad de datos

Hasta aquí, todos los pasos descritos anteriormente deben realizarse independientemente de la técnica que se escoja para la construcción del modelo.

Selección del modelo

La modelización busca la mejor relación de covariables X_i que expliquen la variable dependiente e identificar la ecuación que mejor represente al caso de estudio. Para la construcción del *Modelo matemático* de interés para nuestro caso y escoger el algoritmo del scoring que permita estimar la probabilidad de incumplimiento y el comportamiento de pago de la deuda, lo hicimos de dos maneras:

- Selección del modelo con técnicas tradicionales
- Selección del modelo con técnicas de Inteligencia Artificial

5.1.4.1 Selección del modelo con técnicas tradicionales

El modelo generado fue del tipo *Modelos de regresión de respuesta cualitativa* para nuestro caso *Regresión Logística*, que es una técnica estadística para clasificar los registros a partir de los valores de los campos de entrada. Es análoga a la Regresión Lineal, pero utiliza un campo objetivo categórico en lugar de uno numérico. Se admiten tanto los modelos binomiales (para objetivos con dos categorías discretas) como los multinomiales (para objetivos con más de dos categorías).

La *Regresión Logística* trabaja creando un conjunto de ecuaciones que relacionan los valores de los campos de entrada con las probabilidades asociadas a cada una de las categorías de los campos de salida. Una vez que se ha generado el modelo, se puede utilizar para calcular las probabilidades de datos nuevos. Para cada registro, se calcula una probabilidad de pertenencia a cada categoría posible de salida. La categoría objetivo con la probabilidad más alta se asigna como el valor de salida predicho para cada registro.

El *Nugget de modelo Logístico* contiene la ecuación obtenida para la *Regresión Logística,* en este nodo esta toda la información referente a este modelo, su estructura, parámetros, métricas, etc.

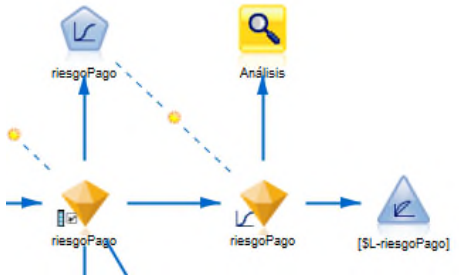


Figura 31: RL - Configuración Nugget (riesgoPago)

En la Figura 31, se observa el Nugget *riesgoPago,* que es un Nugget de modelo Logístico generado a partir del nodo *riesgoPago*, que se encuentra a la izquierda de este nodo, esta Nugget añadió al conjunto de variables seleccionadas para la construcción del modelo dos nuevos campos que contienen la predicción del modelo y la probabilidad asociada. Los nombres de los nuevos campos se derivan del nombre del campo de salida que se está prediciendo, con el prefijo *$L-* para la categoría predicha y *$LP-* para la probabilidad asociada.

Las opciones que se utilizaron para optimizar la generación de los modelos se muestran *en la* Figura 32.

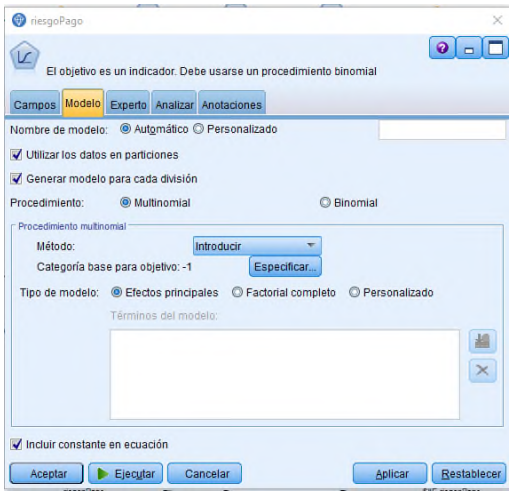


Figura 32: RL - Configuración variable riesgoPago

Características del modelo

Al revisar las características del modelo, se constató que el modelo depende de una sola variable cuantitativa (*plazoOperacion*) como lo muestra la Figura 33 en la pestaña *Modelo,* y que prácticamente no toma en cuenta al resto de variables y que, además, las características del objeto financiado no resultan ser de importancia para el modelo.

Al hacer una revisión más detallada sobre esta variable se observó que los créditos que más problemas tuvieron para el pago y finalmente llegaron al incumplimiento son aquellos que tenían un mayor plazo, lo cual está relacionado directamente con los ingresos, a

menor ventana de ingresos disponibles la propuesta escogida por los clientes era mayor cantidad de cuotas,

En otras palabras esta es una variable que puede ser manipulada aumentando en número de cuotas, y es una variable que las técnicas tradicionales evalúan significativamente, mientras que en ese mismo análisis se observó que los créditos de 24 y 36 meses tuvieron menos problemas cuando se trataba de créditos que tenían objetivos productivos y características de autos que tenían relación con la actividad del cliente, pero esas fueron variables que el modelo de regresión logística no considera importantes.

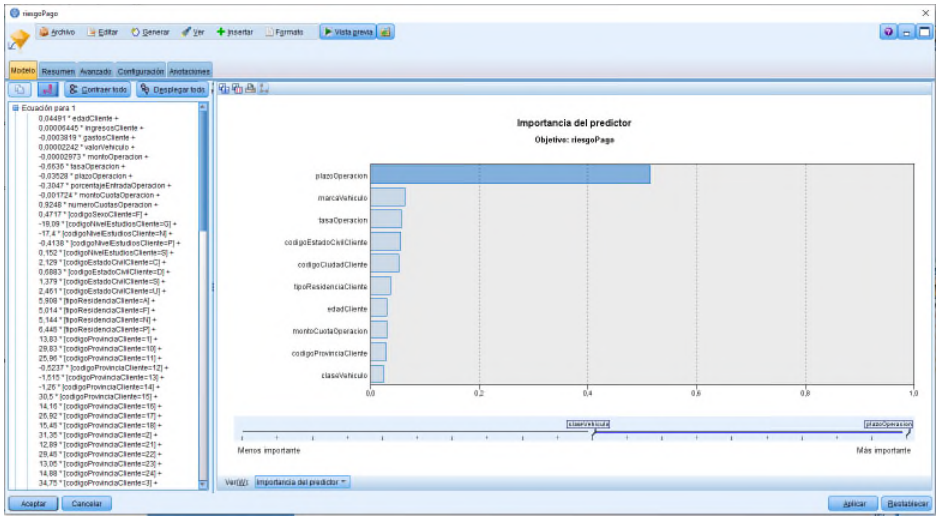


Figura 33: RL - Importancia del predictor

En la Figura 34 en la pestaña *Resumen,* se muestra un resumen del modelo de Regresión Logística.

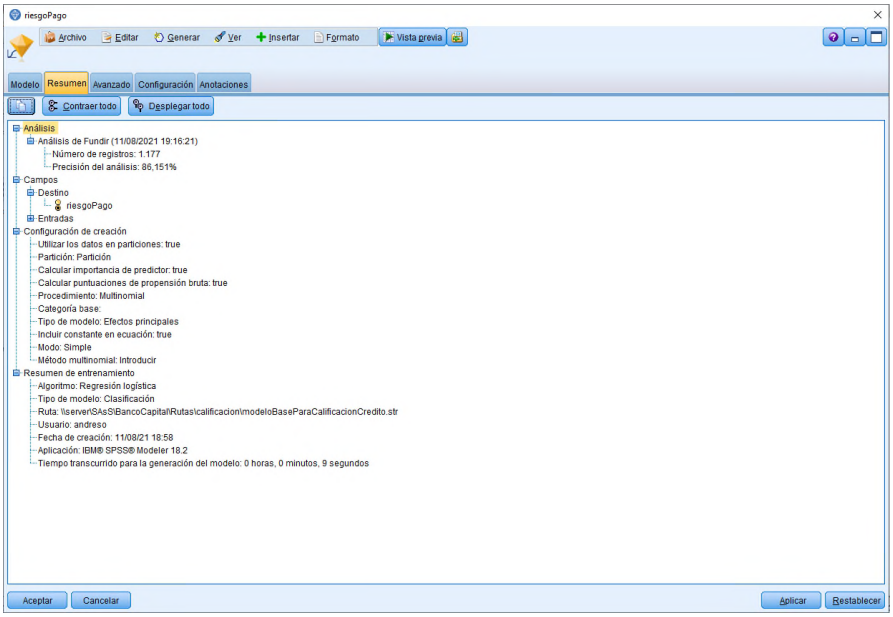


Figura 34: RL - Resumen del modelo

En este tipo de técnicas es común que se durante el desarrollo del modelo se presenten advertencias que indican que los datos tienen problemas, de demuestran la falta de variables para que el modelo sea consistente, ejemplo de ello se muestra en la Figura 35

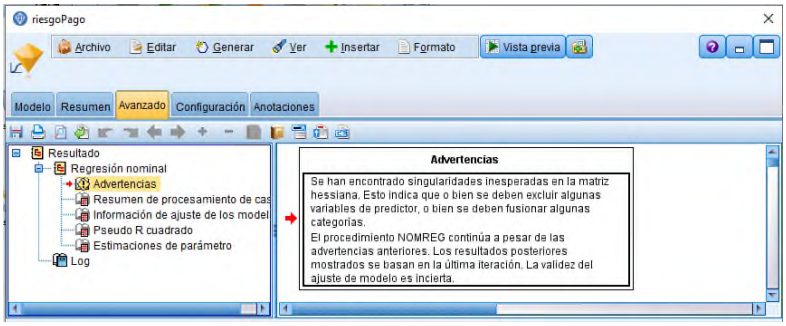


Figura 35: RL - Advertencias

En el reporte mostrado en la Figura 36 también visualizamos que los parámetros del Chi-cuadrado son relativamente muy buenos, pero los Pseudo R cuadrado no lo son.

Información de ajuste de los modelos

Modelo	Criterios de ajuste de modelo	Pruebas de la razón de verosimilitud		
	Logaritmo de la verosimilitud -2	Chi-cuadrado	gl	Sig.
Sólo intersección	1629,628			
Final	715,978	913,650	116	,000

Pseudo R cuadrado

Cox y Snell	,540
Nagelkerke	,720
McFadden	,561

Figura 36: RL - Información de ajuste del modelo

Las características de este modelo se encuentran en el **Anexo *3*:** Características de la Regresión Logística para créditos de consumo.

Validación del modelo diseñado con técnicas tradicionales

Para la regresión logística, después de analizar las variables, observar las correlaciones y de elaborar el modelo de score; fue necesario verificar su fortaleza con base a los métodos estadísticos como: el *R* y *R*2, ANOVA cuyo grado de significancia debe ser igual a cero y el nivel de error o residual debe ser mínimo. El detalle de estas pruebas está definido en ***Validación de los modeles construidos con técnicas estadísticas tradicionales***.

Tabla de comparación del valor observado con el pronosticado

Además de las pruebas, se construyó la tabla de comparación del valor observado con el pronosticado para la variable Riesgo de Pago, donde se registró el número de aciertos para cada una de las categorías de la variable objetivo, en comparación con las equivocaciones para todos los casos.

Esta tabla fue construida teniendo en cuenta que para el entrenamiento del modelo se lo realizó con una porción de los datos y se dejó otra porción para la validación del modelo.

La precisión del modelo la medimos contrastando el número de aciertos con el número de equivocaciones, pero la confianza del modelo se confirma cuando los valores obtenidos en la comparación son similares tanto en el entrenamiento como en la evaluación.

Dicho de otra manera, la confianza en el modelo radica en que se equivoque de la misma manera tanto en el entrenamiento como en la evaluación, esto nos da confianza para cuando se presenten nuevos casos para ser evaluados, el modelo va a tener el mismo número de equivocaciones que obtuvo en el momento del entrenamiento y la evaluación.

La precisión del modelo es importante, pero de nada sirve una muy buena precisión cuando los resultados que se pueden obtener en producción, es decir con nuevos casos, no sean confiables.

En el nodo *Análisis* cuyo resultado de observa en la Figura 37, se evaluó la capacidad de del modelo para generar predicciones precisas. El nodo *Análisis* realiza varias comparaciones entre los valores predichos y los valores reales para uno o más Nuggets de modelo. Los nodos Análisis también se pueden utilizar para comparar modelos.

Nota: Puesto que los nodos Análisis comparan valores predichos con valores reales, solo son útiles con modelos supervisados. Para los modelos sin supervisar, como los algoritmos de agrupación, no existen resultados reales disponibles como base de comparación.

Resultados para el campo de resultado riesgoPago

Modelos individuales

Comparando $L-riesgoPago con riesgoPago

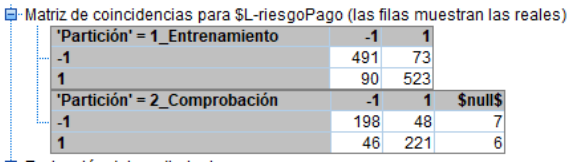
'Partición'	1_Entrenamiento		2_Comprobación	
Correctos	1.014	86,15%	419	79,66%
Erróneos	163	13,85%	107	20,34%
Total	1.177		526	

Figura 37: RL - Resultado, Comprobación

En la tabla comparación de $L-riesgoPago (Pronosticado) con riesgoPago (Observado), se observa que los resultados obtenidos en el entrenamiento son bastante buenos, 86,15% de resultados correctos para este conjunto de datos, pero que en la comprobación 79,66%, además de que disminuye el porcentaje de acierto, difiere mucho del de entrenamiento. Como podemos observar en esta matriz, los resultados para entrenamiento y comprobación son muy diferentes, lo cual genera desconfianza en el modelo. En otras palabras, significa que el modelo no va a poder predecir con la misma precisión a partir de los nuevos casos que se presenten.

Matrices de coincidencias (para objetivos simbólicos o categóricos)

En la Figura 38 se muestra el patrón de coincidencias resultante entre cada categoría de la variable objetivo y su predicción. Se muestra una tabla con filas definidas por valores reales y columnas definidas por valores predichos, con el número de registros que tienen ese patrón en cada casilla. Esto es útil para identificar errores sistemáticos en las predicciones, los casos en los que estos campos concuerdan y no concuerdan se cuentan y se muestran los totales.



Matriz de coincidencias para $L-riesgoPago (las filas muestran las reales)

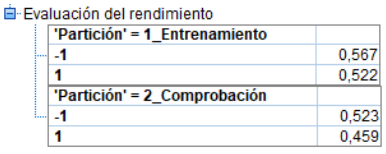
'Partición' = 1_Entrenamiento	-1	1	
-1	491	73	
1	90	523	
'Partición' = 2_Comprobación	-1	1	$null$
-1	198	48	7
1	46	221	6

Figura 38: RL - Resultado, Matriz de coincidencias

Además de lo expuesto anteriormente en la matriz de comparación, se observa en esta tabla en mayor detalle la clasificación de buenos y malos créditos, para este caso se reafirma lo expresado anteriormente. Tal como hemos definido anteriormente, los valores de -1 se usa para créditos malos y 1 para créditos buenos. El modelo tiene aparentemente similar acierto para los créditos buenos (85%) que para los malos (87%) y eso parecería ser muy bueno, pero en la etapa de comprobación no se obtiene los mismos porcentajes de acierto, (31) para los buenos y (80%) para los malos y esto genera desconfianza.

Evaluación del rendimiento

Uno de los resultados que se obtiene se muestra en la Figura 39, que es un conjunto de valores estadísticos que dan la medida del rendimiento para modelos con resultados categóricos. Estos valores estadísticos se muestran para cada categoría de los campos de salida y consiste en una medida del contenido de información medio (en bits) del modelo, para predecir registros pertenecientes a cada categoría. Se tiene en cuenta la dificultad del problema de clasificación, de forma que las predicciones precisas para categorías inusuales obtendrán un índice de evaluación del rendimiento mayor que las predicciones precisas para categorías comunes. Si el modelo no es asertivo en el valor real de una categoría, el índice de evaluación del rendimiento para esa categoría será tendiente a cero.



Evaluación del rendimiento

'Partición' = 1_Entrenamiento	
-1	0,567
1	0,522
'Partición' = 2_Comprobación	
-1	0,523
1	0,459

Figura 39: RL - Resultado, Evaluación del rendimiento

Los resultados no son buenos.

Métricas de evaluación (AUC & Gini, solo para clasificadores binarios)

En el caso de los clasificadores binarios, se utiliza las métricas de evaluación de coeficiente Gini y AUC (Area Under Curve). Ambas métricas se calculan de forma conjunta para cada modelo binario. Los valores de las métricas se muestran en la Figura 40.

Métricas de evaluación

'Partición'	1_Entrenamiento		2_Comprobación	
Modelo	AUC	Gini	AUC	Gini
$L-riesgoPago	0,942	0,884	0,866	0,733

Figura 40: RL - Resultado, Métricas de evaluación

La métrica de evaluación AUC se calcula como el área bajo la curva ROC (Receiver Operator Characteristic) y es una representación escalar del rendimiento esperado de un clasificador. El AUC se sitúa siempre entre 0 y 1, y cuanto más alto es el valor, mejor es el clasificador. Una curva ROC diagonal entre las coordenadas (0,0) y (1,1) representa un clasificador aleatorio y tiene un AUC de 0,5. Así pues, un clasificador realista no tendrá un AUC de menos de 0,5.

La métrica de evaluación de coeficiente Gini se utiliza a veces como métrica de evaluación alternativa a la AUC, y ambas medidas están estrechamente relacionadas. El coeficiente Gini se calcula como dos veces el área comprendida entre la curva ROC y la diagonal, o como Gini = 2AUC – 1. El coeficiente Gini está siempre entre 0 y 1, y cuanto mayor es el número, mejor es el clasificador. El coeficiente Gini será negativo en el improbable caso de que la curva ROC esté por debajo de la diagonal.

Tal como se describe en la definición de los parámetros, los valores son muy superiores a 0 y parecerían ser bastante buenos, pero los resultados no son confiables, ya que los mismos varían significativamente en cada una de las muestras.

Evaluación SFX RiesgoPago

Figura 41: RL - Ejecución SFX RiesgoPago

El nodo Evaluación mostrado en la Figura 41 ayuda a evaluar y comparar modelos predictivos. El diagrama de evaluación de las Figura 42 y Figura 43 muestran la calidad con que los modelos predicen resultados particulares. Ordena registros en función del valor predicho y la confianza de la predicción. Divide el registro en (**cuantiles**) y, a

continuación, representa el valor del criterio de negocio de cada cuantil de mayor a menor. El gráfico muestra múltiples modelos como líneas independientes.

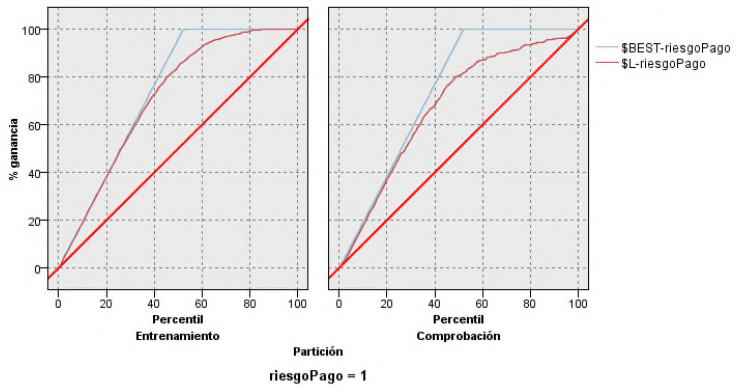


Figura 42: RL - Resultado, Evaluación y comparación, Créditos buenos

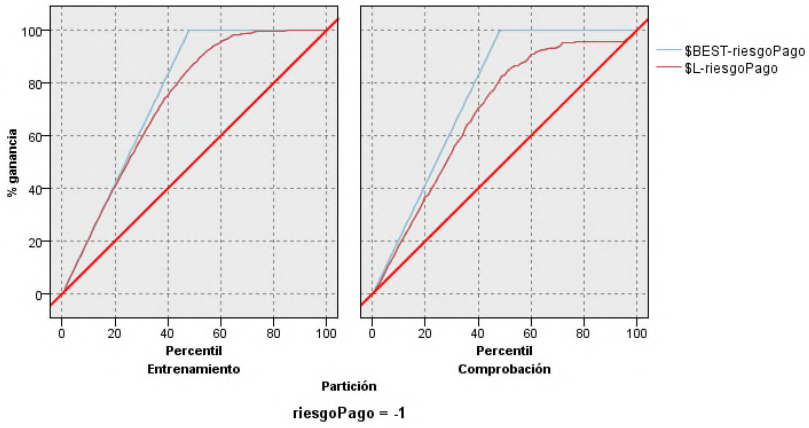


Figura 43: RL - Resultado, Evaluación y comparación, Créditos malos

Tal como se puede observar en la Figura 42 y Figura 43, el área bajo la curva no es similar tanto para el entrenamiento, como para la comprobación, aunque el área cubierta es bastante cercana a la ideal.

5.1.4.2 Selección del modelo con técnicas de Inteligencia Artificial

Muchas veces, las condiciones que se han dado para el pago o no de un crédito no son bien comprendidas por los modelos tradicionales, y es ahí cuando los modelos creados con tecnologías de inteligencia artificial resultan ser más adecuados, estos incorporan en su entrenamiento conocimiento que luego les permite calificar una operación de crédito de maneras más precisas y variadas.

Con el Modelado Automático en este caso el llamado *Clasificador automático,* como el que se muestra en la Figura 44, se estimaron los modelos candidatos para cada combinación de opciones posible. En el Nugget de modelo resultante se integraron los

modelos que mejor se ajustaron a los datos formando un único Nugget de modelo automático compuesto.

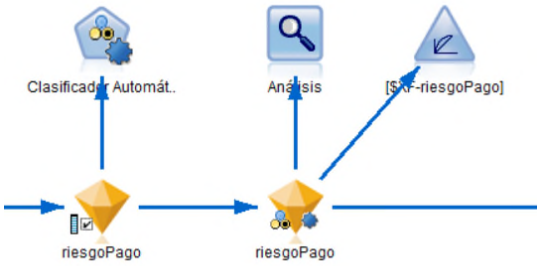


Figura 44: IA - Configuración Nodo Clasificador Automático

En la Figura 45, se muestra en el nodo *Clasificador Automático*, como quedaron configuradas las opciones que se utilizaron para optimizar la generación de los modelos.

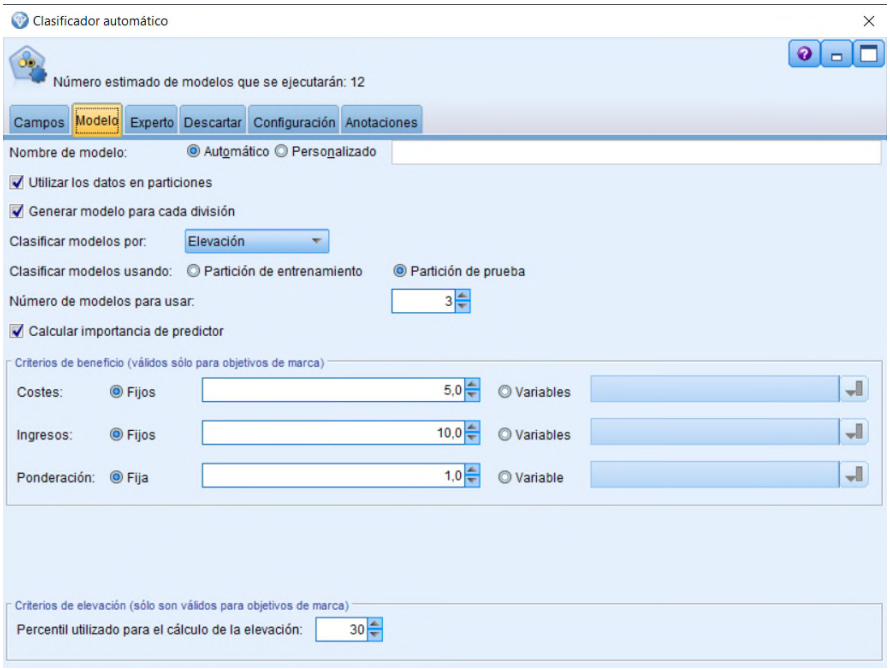


Figura 45: IA - Configuración modelo

En la Figura 46, en la pestaña Experto, se seleccionó cuál de las varias técnicas existentes en la herramienta se utilizarán para comparar los resultados. Este nodo genera un conjunto de modelos basado en las opciones dadas y ranquea a los mejores candidatos.

Nótese que se han seleccionado técnicas tradicionales como: Regresión Logística, Discriminante y no tradicionales como: Árboles de Decisión, redes Neuronales, entre otras.

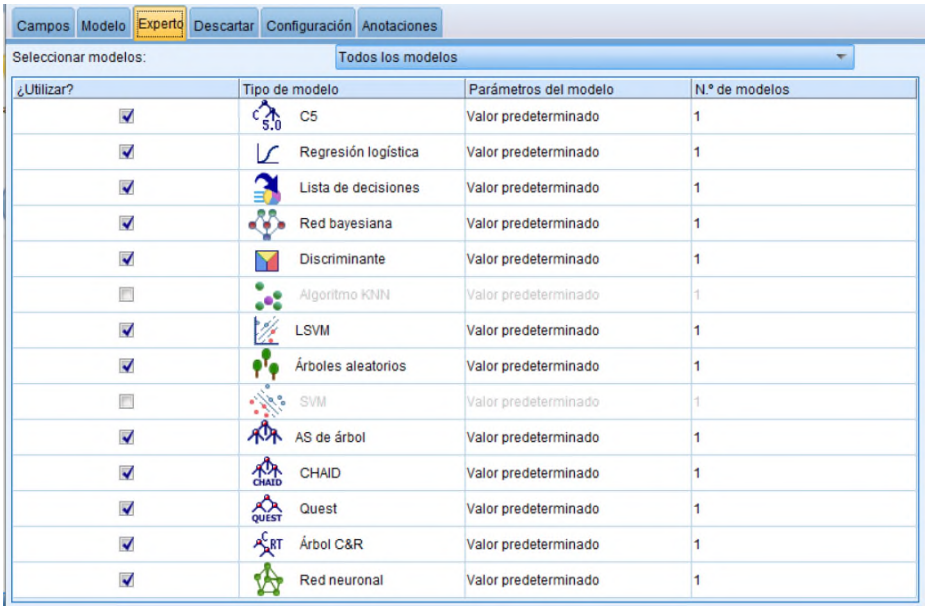


Figura 46: IA – Selección de técnicas

Diseño del modelo

Los modelos seleccionados, los cuales fueron examinados individualmente se muestran en la Figura 47.

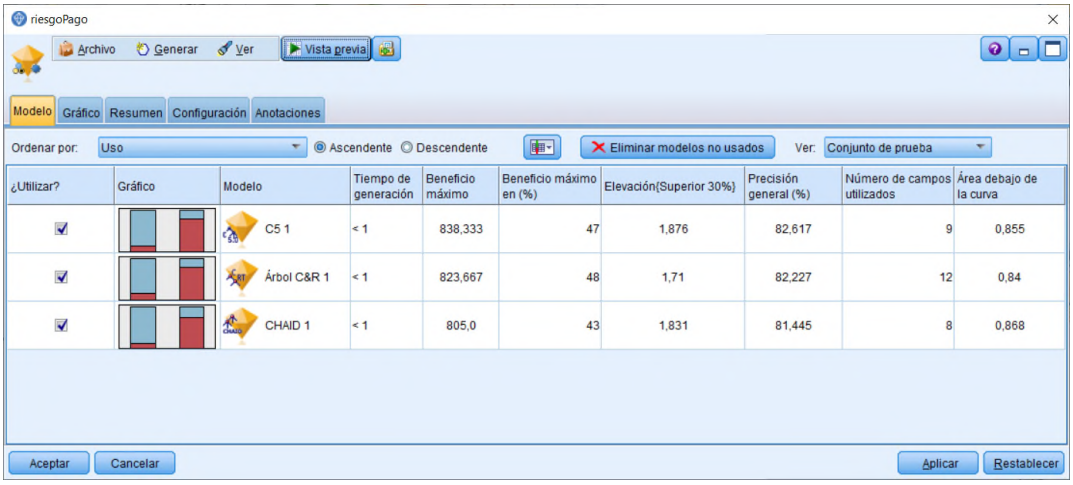


Figura 47: IA - Resultado, Modelos generados

El número de modelos seleccionados es impar y mayor a 3, para que exista mayoría en la votación en caso de que dos modelos clasifiquen de manera contraria. Los modelos escogidos son: C5_1, Árbol de Clasificación y Regresión_1 y CHAID_1, donde todos ellos pertenecen a la técnica de modelado Árboles de Clasificación (Inteligencia Artificial).

En la pestaña Gráficos de la Figura 48, se observa de manera general la clasificación que hacen los modelos seleccionados (Gráfico de barras – Distribución), en el cual se observa que hay mejor aproximación para los casos malos que para los buenos. En el panel

derecho se encuentra el gráfico *Importancia del predictor*, el mismo que indica el grado de aportación de cada variable con respecto a la variable *Destino*.

Figura 48: IA - Resultado, gráfico de modelos generados

Como podemos observar en la Figura 48, el resultado final en cuanto a la importancia del predictor fue bastante bueno ya que prácticamente les da la misma importancia a 10 variables que, a diferencia del modelo anterior, solo se la da a una variable.

Para nuestro caso y en favor de nuestro proyecto incluye dos variables que hacen referencia al objeto financiado: la marca y la clase de vehículo, ambas variables cualitativas. Estas variables demostraron ser muy aportantes en la definición del modelo, sobre todo en la determinación de los casos malos.

Las características de cada modelo se encuentran en el **Anexo *4*:** Características del modelo con IA para créditos de consumo.

A continuación, se muestran los resultados de la validación del modelo.

Validación del modelo deseñado con IA

Para el caso que el modelo seleccionado sea construido con Inteligencia Artificial la validación del modelo resulta mucho más simple, pues es suficiente con las tablas que a continuación se muestran en donde se registra el número de aciertos y equivocaciones para cada una de las categorías de la variable objetivo.

Tabla de comparación del valor observado con el pronosticado

Resultados para el campo de resultado riesgoPago

Modelos individuales

Comparando $XF-riesgoPago con riesgoPago

'Partición'	1_Entrenamiento		2_Comprobación	
Correctos	960	83,99%	429	83,95%
Erróneos	183	16,01%	82	16,05%
Total	1.143		511	

Figura 49: IA - Resultado, Tabla de comprobación

En la Figura 49 se muestra el resumen de los resultados de esta tabla.

En la tabla comparación de $XF-riesgoPago (Pronosticado) con riesgoPago (Observado), se observa que los resultados obtenidos en cada una de las partes son muy buenos, más o menos 84% de resultados correctos, esto se refiere a la precisión del modelo. Pero lo más importante en esta tabla, es que estos resultados se conservan para cada una de las etapas de la construcción del modelo, lo cual genera confianza en el mismo. En otras palabras, significa que el modelo va a poder predecir con la misma precisión a partir de los nuevos casos que se presenten.

Matriz de coincidencia

La Figura 50, se muestra el patrón de coincidencias entre cada categoría de la variable objetivo y su predicción.

Matriz de coincidencias para $XF-riesgoPago (las filas muestran las reales)

'Partición' = 1_Entrenamiento	-1	1
-1	512	53
1	130	448
'Partición' = 2_Comprobación	-1	1
-1	227	25
1	57	202

Figura 50: IA - Resultado, matriz de coincidencias

Además de lo expuesto anteriormente, en la matriz de coincidencias, se observa en esta tabla en mayor detalle la clasificación para buenos y malos, para este caso se reafirma lo expresado anteriormente. Tal como hemos definido anteriormente, los valores de -1 son para créditos malos y 1 para créditos buenos. El modelo tiene mayor acierto en los malos (91%), para los buenos (78%). Lo más importante es que en la fase de prueba se obtiene porcentajes de acierto muy similares, para los créditos malos y buenos, 90% y 78% respectivamente.

Evaluación del rendimiento

En la Figura 51 se muestra el estadístico de evaluación del rendimiento para modelos con resultados categóricos.

Evaluación del rendimiento

'Partición' = 1_Entrenamiento	
-1	0,478
1	0,57
'Partición' = 2_Comprobación	
-1	0,483
1	0,563

Figura 51: IA - Resultado, evaluación del rendimiento

Los resultados de estos parámetros no son buenos.

Métricas de evaluación (AUC & Gini, solo para clasificadores binarios)

En el caso de los clasificadores binarios, se utiliza las métricas de evaluación de coeficiente Gini y AUC (Area Under Curve). Ambas métricas se calculan de forma conjunta para cada modelo binario. Los valores de las métricas se ilustran en la Figura 52.

Métricas de evaluación

'Partición'	1_Entrenamiento		2_Comprobación	
Modelo	AUC	Gini	AUC	Gini
$XF-riesgoPago	0,903	0,805	0,909	0,819

Figura 52: IA - Resultado, métricas de evaluación

Los diagramas de evaluación mostrados en las Figura 53 y Figura 54 muestran la calidad con que los modelos predicen resultados particulares. El gráfico muestra múltiples modelos como líneas independientes.

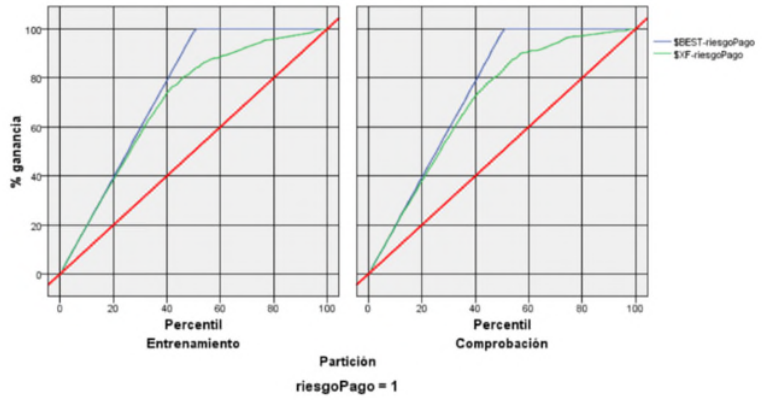


Figura 53: IA - Resultado, evaluación y comparación, créditos buenos

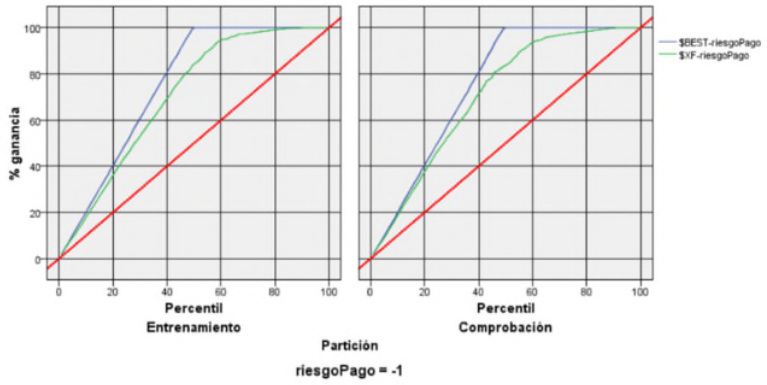


Figura 54: IA - Resultado, evaluación y comparación, créditos malos

Como se puede observar en ambos gráficos, el área bajo la curva es similar tanto para el entrenamiento, como para la comprobación y el área cubierta es bastante cercana a la ideal.

5.2 Modelo para el financiamiento Automotriz - Microcrédito

A continuación, detallamos las etapas para el desarrollo del modelo para otorgamiento de un microcrédito destinado al financiamiento automotriz.

5.2.1 Entendimiento del negocio

Microcrédito es el préstamo que concede una institución financiera y que se destina a financiar actividades de producción y/o comercialización en pequeña escala, cuya fuente principal de pago la constituye el producto de las ventas o ingresos generados por dichas actividades.

Se supone que para su otorgamiento y para la determinación del monto, cuotas, etc., se analizan también las ventas o ingresos que adicionalmente se generaran en la actividad comercial con la incorporación del bien financiado, pero en realidad, en la mayoría de los casos esto no ocurre así, principalmente debido a la dificultad que implica hacer ese estudio previo para cada una de las solicitudes.

Finalmente, de igual manera que para el caso de créditos de consumo, casi siempre los créditos son otorgados por la determinación de la capacidad de pago, que depende del nivel de ingresos del prospecto, menos los gastos de la actividad comercial y menos los pagos a entidades financieras.

Igual que en los créditos de consumo en este tipo de créditos, la contribución de un modelo clásico, construido con técnicas tradicionales es bastante limitada, por un lado, porque si existe disponibilidad para el pago del crédito el modelo seguramente aprobará la operación y, por otro lado si el prospecto no tiene capacidad de pago o no puede justificar los ingresos adicionales que obtendrá con la adquisición, pero ha tenido un buen récord crediticio, un modelo como los anteriores no podrá ayudarlo mucho y lo hará aún menos si se trata de un modelo diseñado con técnicas tradicionales.

De igual manera esto no generará inclusión financiera, por lo que a el modelo propuesto pretende mejorar esta situación, pues incluye variables que pretenden incorporar el aporte que tendría el bien a financiar en los ingresos adicionales, las cuales se espera resulten importantes en la determinación de la probabilidad de pago,

En el Ecuador la definición de este tipo de créditos está dada en el glosario de términos en la parte correspondiente a ***Microcrédito***.

Objetivo

Orientar la metodología CRIPS-DM para seguir los lineamientos definidos en la metodología que aplicará el banco para el otorgamiento, seguimiento y recuperación de microcrédito con base a la normativa, políticas y procesos internos.

Alcance

La metodología definida por el banco contiene un conjunto de métodos, herramientas e instrumentos específicos para la ejecución de las etapas del proceso de crédito, que serán de uso y cumplimiento obligatorio para todas las Agencias, y será aplicado específicamente por los Asesores de Crédito, Gerente de Negocios e integrantes del Comité de Crédito, instancias de control interno, departamento de riesgos y por cualquier empleado que esté relacionado con la administración o gestión del proceso de crédito.

Descripción de la metodología del banco

El Asesor de crédito, es quien realiza todo el proceso metodológico desde el contacto con el cliente, el proceso de evaluación, aprobación, y renovación de la operación crediticia. A continuación, se presenta un resumen de las etapas que conforman la metodología de microcrédito individual.

Prospección: El Asesor de crédito, en esta etapa, se informa sobre las características del cliente, en lo concerniente a la actividad económica y análisis de la competencia; por lo tanto, esta etapa proporciona el conocimiento previo del mercado objetivo.

Promoción: Es una tarea permanente que realizan los Asesores de crédito mediante campañas de promoción del producto a clientes pertenecientes al mercado objetivo. La promoción siempre debe cumplir las siguientes premisas:

- Información sobre las condiciones del producto como tasa, plazo, interés, etc.
- Beneficios del producto
- Proporcionar la información sobre los requerimientos con que el cliente debe contar al momento de solicitar el crédito.
- Material publicitario con información del producto y datos de contacto del Asesor de crédito.

Inspección de campo: Una vez que el potencial cliente demuestra interés y entrega la documentación solicitada, el asesor de crédito acude al levantamiento de información financiera. Esta visita es importante y determinante, pues de ella depende la calificación y posterior aprobación de crédito.

La inspección de campo constituye un elemento fundamental para la evaluación crediticia porque se registra la percepción del asesor, a través de los siguientes aspectos:

- **La presentación del negocio:** Se califica la organización, orden y aseo del lugar, son factores que pueden determinar la aprobación o negación del crédito.
- **El nivel de inventarios:** Sean en materia prima, en proceso o productos terminados, artículos que posee el sujeto de crédito al momento del levantamiento de la información.

- **El nivel de actividad:** Además de la presentación se mide la concurrencia de los clientes al negocio.
- **La ubicación:** Se determina si el negocio está ubicado en un lugar estratégico o comercial, caso contrario puede existir un alto riesgo el otorgar un crédito a un negocio que tiene una mala ubicación.
- **La percepción de los clientes:** Durante la inspección del asesor de crédito, se realizan preguntas clave a los clientes, para confirmar la veracidad de los datos aportados por el sujeto de crédito: ¿Quién es el propietario del negocio?, ¿Cómo ve usted el negocio?, ¿Cómo es el dueño del negocio?, ¿Le considera una persona responsable?, son preguntas que se las hace para que el asesor pueda definir si es una persona que dijo o no la verdad, pudiendo ser un punto de apoyo en la decisión de otorgar el crédito.
- **Percepción de la competencia:** Al mismo tiempo que se mide la percepción de los vecinos se puede también ver el nivel de competencia que existe alrededor del negocio, con esto se puede evidenciar si existe o no competencia para ese negocio.

5.2.2 Entendimiento de los datos

Luego del levantamiento de la información, se procedió a la evaluación de ésta; entre los factores que se consideran para la evaluación financiera del potencial cliente se toma en cuenta la solvencia, considerando las 5 C del crédito como son: carácter, capacidad, capital, condiciones y colateral.

Carácter: Se refiere a conocer al solicitante; su historial crediticio con el banco y otras instituciones, referencias comerciales y el entorno. Con este conocimiento se tiene una primera apreciación de su carácter.

Es un análisis inicial referencial del comportamiento del solicitante (basado en perfiles predefinidos), esta información se obtiene de registros del banco o en la visita al negocio, mediante referencias del entorno en donde está ubicado el negocio o el hogar del solicitante. La información relevante es:

- **Características generales del negocio:** Tiempo de permanencia, experiencia en la actividad, condiciones del mercado, entorno del local y que la actividad tenga las condicionantes requeridas por el mercado para asegurar su permanencia en el largo plazo.
- **Referencias (comerciales y personales):** para medir el comportamiento de los solicitantes con alguna empresa o proveedor donde adquiere regularmente sus productos y también las referencias personales. Ambas proporcionarán datos importantes en cuanto a su comportamiento y la forma en que maneja sus obligaciones.

- **Ubicación y permanencia del negocio:** por lo general un negocio que muestra experiencia y además se encuentra en lugares de alta concentración, buena ubicación, afluencia de clientes, son elementos diferenciadores entre negocios "de alto y bajo riesgo". La estabilidad del negocio y experiencia del propietario será un factor determinante en la decisión del asesor de crédito.

Capacidad: Permite levantar y cruzar información relevante como:

- Se evaluará al endeudamiento de consumo y su capacidad de afrontar las deudas con sus ingresos.
- La experiencia en la actividad económica
- La organización del negocio
- La dependencia que tenga de los clientes
- La dependencia de los proveedores
- La dependencia del personal, si lo tuviera
- El nivel de formación del propietario del negocio

En la determinación de la capacidad se debe incluir el aporte dado por el objeto a financiar, es decir la contribución que tendrá en el negocio el destino del crédito.

El capital se considerará por**:**

- La capacidad de pago, de acuerdo con la inspección y respaldos
- La rentabilidad que obtenga por la venta de sus productos o servicios
- El nivel de endeudamiento

Las condiciones determinan el comportamiento del negocio o empresa del sujeto de crédito, que incluye:

- El nivel de la competencia del sector
- La aceptación del producto que ofrece al cliente

El colateral estará medido por: las garantías o apoyos colaterales que se evalúan a través de sus activos fijos, el valor económico y calidad de estos, se puede tomar para compensar una debilidad de las principales C de crédito, pero nunca sustituir al carácter.

- **La garantía ofrecida.**

El procedimiento de calificación de los factores que se consideran para la evaluación financiera del potencial cliente, que en definitiva es la solvencia, comprende el determinar la condición que presenta o refleja el negocio en el momento de la solicitud de crédito; cada criterio dispone de diferentes niveles valorados por una puntuación que están parametrizados con anterioridad en la política de crédito del banco.

La descripción detallada de este procedimiento se encuentra en el **Anexo *2*:** Procedimiento de calificación para la evaluación financiera.

A continuación, se realizó a manera de experimentación, la ponderación del prospecto basada en los criterios antes mencionados, sobre todo para confirmar el aporte de estas nuevas variables.

La determinación de la ponderación se la hace en base a la importancia que tuvieron estos criterios en un modelo donde participaron exclusivamente estas variables para la determinación de la probabilidad de pago. Los resultados obtenidos se muestran gráficamente en la Figura 55 y cuantificadamente en la Tabla 7.

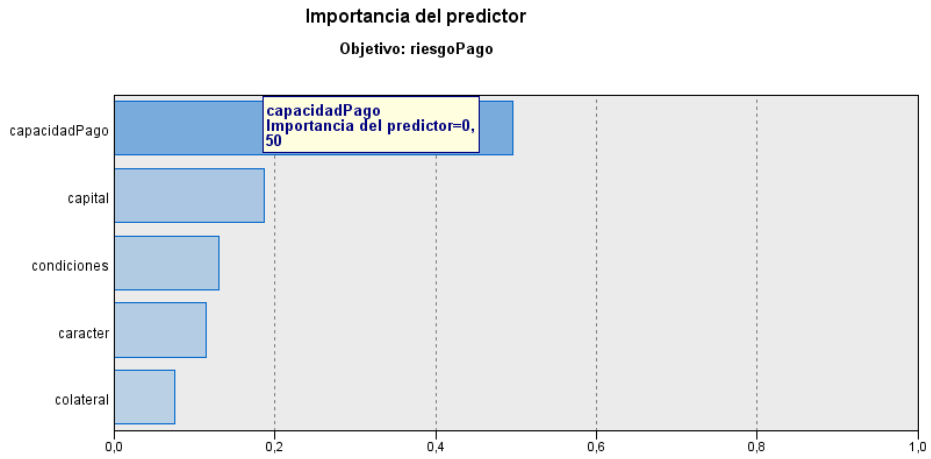


Figura 55: Importancia de los criterios para la evaluación financiera

Tabla 7: Resumen de la importancia de los criterios en la evaluación financiera

VARIABLE	PESO
CARÁCTER	11%
CAPACIDAD DE PAGO	50%
CAPITAL	19%
CONDICIONES	13%
COLATERAL	7%
TOTAL	**100%**

Según esta ponderación, el número mínimo de puntos necesario para que una solicitud de crédito sea aprobada, se estima en 70%, ya que según la Figura 56: IA - Score de riesgo de pago Vs puntos obtenidos, es a partir de ese valor donde hay un quiebre importante entre el número de casos donde el riesgo de no pago es significativamente menor, sin embargo, se podrá hacer excepciones que pueden ir desde el 50%.

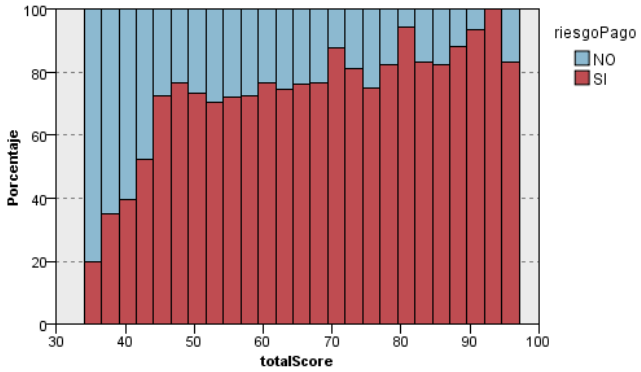


Figura 56: IA - Score de riesgo de pago Vs puntos obtenidos

Calidad del microcrédito

Para asegurar la colocación de un buen crédito, el banco dispone de un modelo de solicitud de crédito, así como un **modelo de score propio** que permite realizar una investigación adecuada, previa a la aprobación y desembolso del crédito. Para el proceso de investigación el cliente debió entregar la documentación que respalda lo declarado en la solicitud.

A continuación, describiremos el desarrollo del modelo de score, el cual incorpora las variables antes descritas en la metodología para el otorgamiento del microcrédito por parte del banco, las cuales fueron tomadas para el otorgamiento de microcrédito automotriz.

A medida que se obtenga datos adicionales en respuesta al pago de los microcréditos, se procederá a reentrenar este modelo y así reformular matemáticamente su aporte en la determinación de la probabilidad de pago.

Finalmente, a los datos anteriormente analizados en esta misma etapa se agregan los siguientes:

Variables relacionadas con la evaluación financiera

Criterio	Nombre de la variable	Nombre en el modelo	Tipo	Descripción
Carácter	Comportamiento crediticio	carac_comportCre	Decimal	Rango Puntuación de 1-10
Carácter	Calificación Buro	carac_buro	Decimal	Rango Puntuación de 1-10
Carácter	Entrevista	carac_entrevista	Decimal	Rango Puntuación de 1-10
Carácter	Zona geográfica	carac_zonaGeo	Decimal	Rango Puntuación de 1-10
Capacidad	Situación financiera	capPago_relMontoCuota	Decimal	Rango Puntuación de 1-10

Capacidad	Destino del crédito	capPago_destinoCred	Decimal	Rango Puntuación de 1-10
Capital	Capital	capital	Decimal	Rango Puntuación de 1-10
Condiciones	Condiciones	condiciones	Decimal	Rango Puntuación de 1-10
Capital Colateral	Capital Colateral	Capital colateral	Decimal Decimal	Rango Puntuación de 1-10 Rango Puntuación de 1-10

5.2.3 Preparación de los datos

Esta fase es similar a la realizada anteriormente para los créditos de consumo hasta la determinación de los casos buenos y malos, luego de lo cual se filtran los casos pertenecientes a la categoría de microcrédito, después de lo cual quedan 2.656 créditos como lo muestra la Figura 57.

	idCliente	idOperacion	tipoOperacionOriginal	capital	condiciones	colateral	caracter	capacidadPago
1	0100994185	LD1207900026	MICROCREDITO	5.000	10.000	10.000	8.886	2.000
2	0101196517	LD1234200030	MICROCREDITO	5.000	2.000	6.000	6.743	6.800
3	0101266716	LD1117100044	MICROCREDITO	5.000	2.000	6.000	8.086	2.400
4	0101647378	LD1216400030	MICROCREDITO	5.000	4.000	8.000	7.486	1.600
5	0101797611	LD1123700032	MICROCREDITO	5.000	10.000	8.000	5.229	10.000
6	0102045390	LD1431500006	MICROCREDITO	5.000	2.000	6.000	6.429	2.000
7	0102074002	LD1503500023	MICROCREDITO	5.000	10.000	10.000	4.200	2.800
8	0102078789	LD1314700065	MICROCREDITO	5.000	2.000	6.000	8.400	2.400

Figura 57: Registros de microcrédito automotriz

5.2.4 Modelado

Equilibrio de la muestra

Tal como lo muestra la Figura 58, se obtuvo la siguiente proporción de datos de cartera considerada buena y mala.

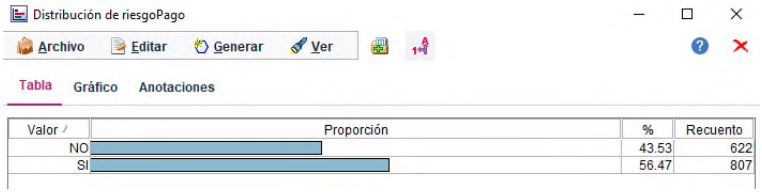


Figura 58: IA (Microcrédito) - Distribución de las categorías de riesgo de pago

Se observa que la muestra quedo bastante pareja siendo una diferencia de 56% a 44% que es considerada pertinente para entrenar al modelo y que pueda aprender en similar proporción de las dos categorías.

Luego de ello se procede a entrenar el modelo con los 1.429 registros de la muestra, tal como se observa en la Figura 59.

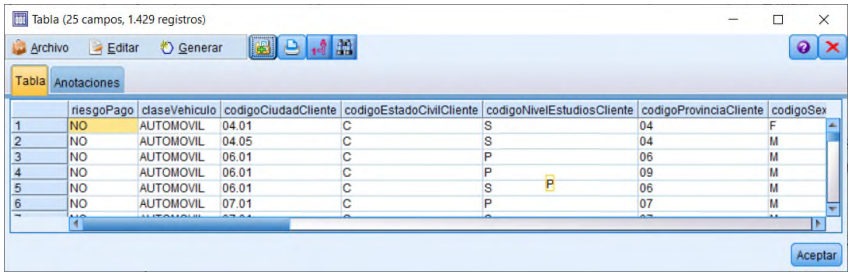


Figura 59: IA (Microcrédito) - Registros para entrenamiento del modelo

Partición de la muestra

Las opciones de edición del nodo *Partición* se realizaron de igual manera que en los modelos anteriores

Selección de variables

En el listado mostrado en la Figura 60 se observan las variables que podrían participar en el modelo.

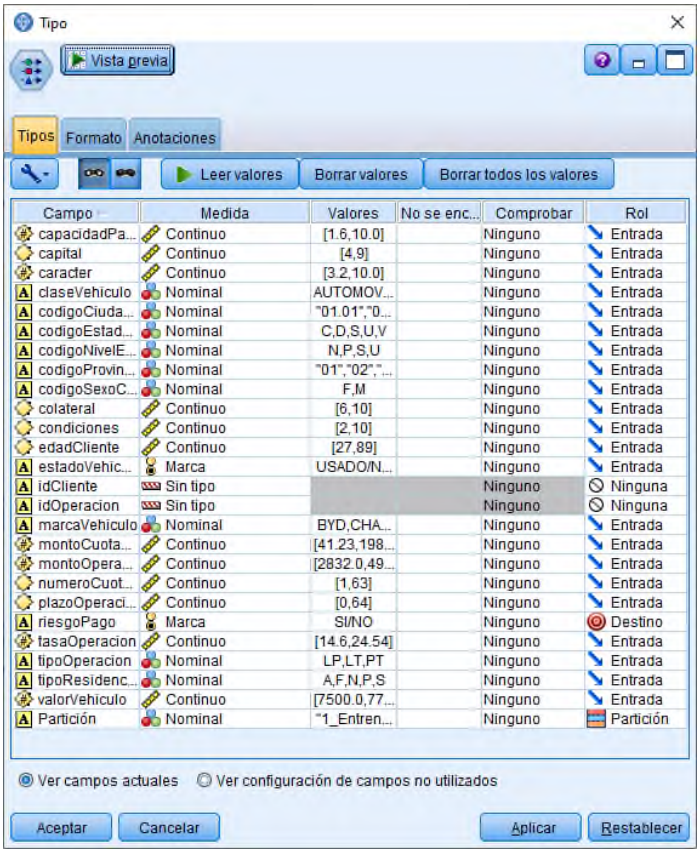


Figura 60: IA (Microcrédito) - Variables que participarían en el modelo

Con el nodo *Selección de Características* se determinó estadísticamente mediante la discriminación de los registros en cada categoría de la variable objetivo, el nivel de aporte que tiene cada variable, el resultado es el mostrado en la Figura 61.

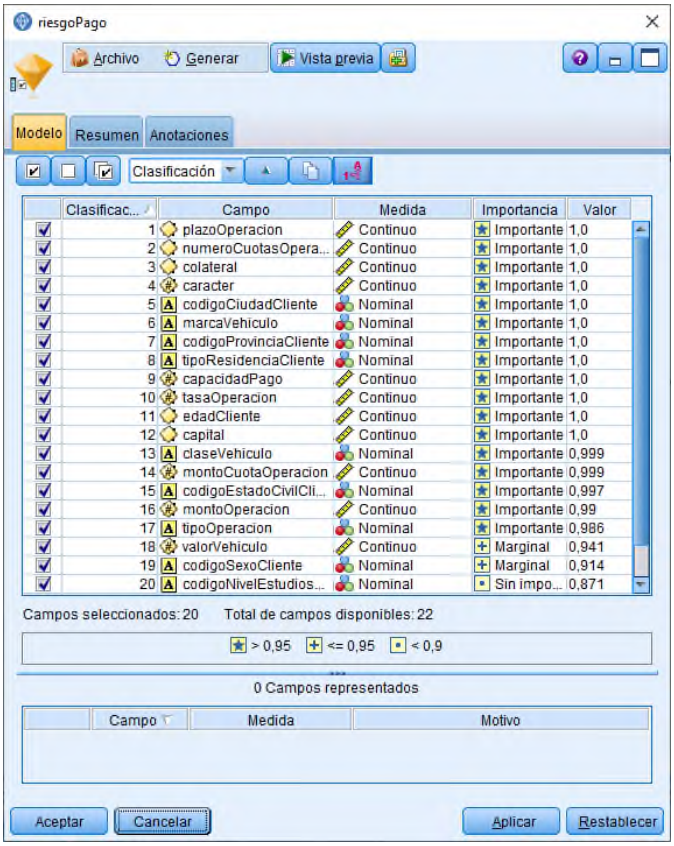


Figura 61: IA (Microcrédito) - Resultado del orden de importancia de variables

En este resultado se observa que las 15 primeras variables son relevantes para el modelo, las variables restantes aportan muy poco. Sin embargo, para posibilitar la integración con el proceso de calificación de crédito, se han incorporado estas 5 variables restantes, las cuales no serán tomadas en cuenta por el modelo, pero al tratarse de un modelo de inteligencia artificial, se espera que, al reentrenarse, estas variables tengan el suficiente coeficiente de variación y aportar al modelo, junto con nuevas variables que irán apareciendo, conforme se acumule más información para la calificación.

Se observó la distribución de cada una de las variables respecto a la variable *Destino*, a fin de evaluar la calidad de los datos; por ejemplo, en la Figura 62 se muestra el resultado para la variable *montoOperación,* que se refiere al monto del crédito.

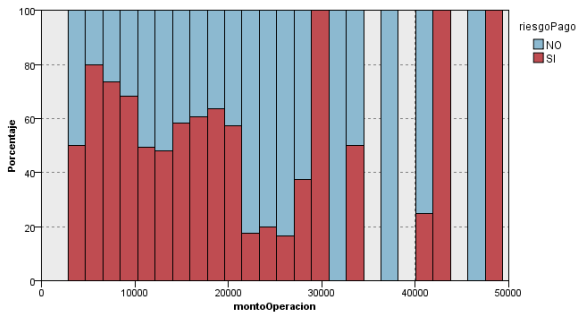


Figura 62: IA (Microcrédito) - Resultado, variable monto de la operación - Normalizada.

En Figura 62 el valor *NO* corresponde a créditos malos y *SI* a créditos buenos. Si la variable no aportara todas las barras tendrán igual altura para cada una de las categorías. Este no es el caso.

De esta forma, se comprobó gráficamente que la mayoría de las variables independientes y en especial las cualitativas aportaban significativamente al modelo tal como lo indicaba el nodo *Selección de Características*.

5.2.4.1 Selección del modelo

Con el nodo *Clasificador Automático* de la Figura 63, se estimaron los modelos candidatos para cada combinación de las posibles opciones. Se guardaron los mejores modelos en un *Nugget* de modelo automático compuesto.

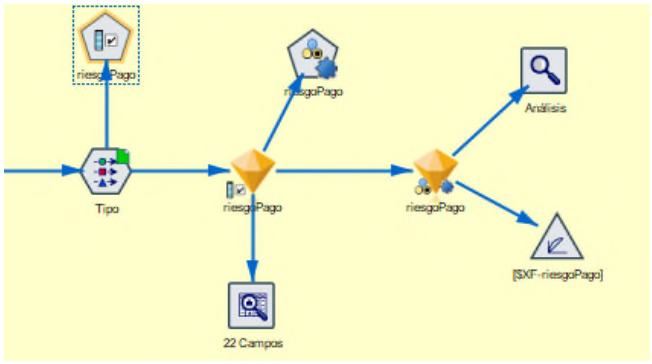


Figura 63: IA (Microcrédito) - Configuración Nodo Clasificador Automático

En la Figura 64, se observa cuáles son las técnicas de la herramienta que se utilizarán para comparar los resultados. Se muestra que se han seleccionado técnicas tradicionales como: Regresión Logística, Discriminante y no tradicionales como: Árboles de Decisión, redes Neuronales, entre otras

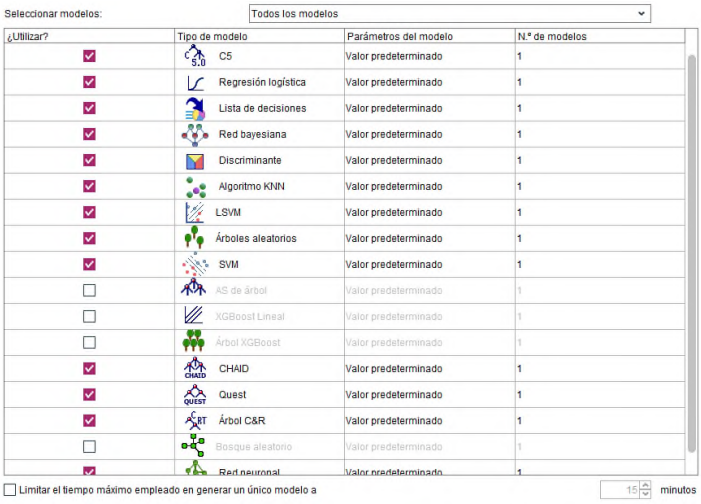


Figura 64: IA (Microcrédito) – Configuración Nodo riesgoPago

Construcción del modelo

El número de modelos seleccionados de la lista mostrada en la Figura 66, siempre será impar y mayor a 3, para que exista mayoría en la votación en caso de que dos modelos clasifiquen de manera contraria. Los modelos escogidos son: *Bosque aleatorio* 1, *Lista de decisiones 1* y *Árbol XGBoost 1*, donde todos ellos pertenecen a la técnica de modelado Árboles de Clasificación (Inteligencia Artificial).

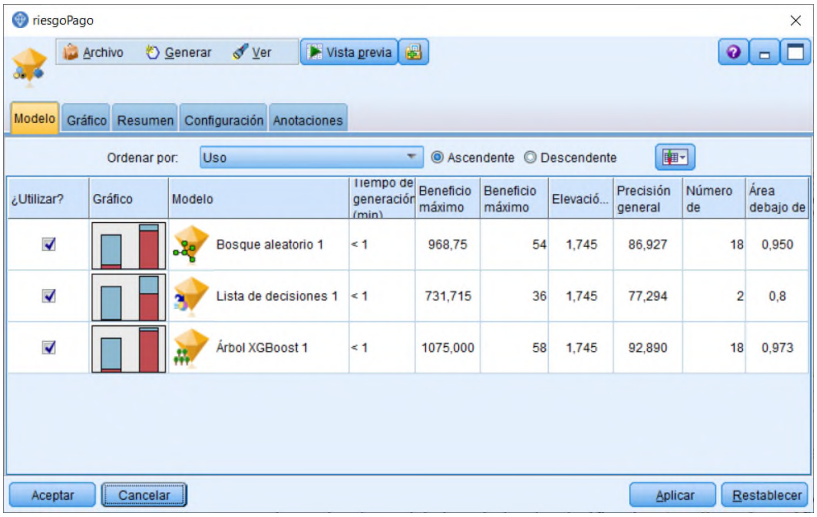


Figura 65: IA (Microcrédito) - Resultado, modelos generados

En la Figura 66, se observa de manera general la clasificación realizada por los modelos escogidos (Gráfico de barras – Distribución), en el cual se observa que hay mejor aproximación para los malos que para los buenos. En el panel de la derecha se encuentra el gráfico *Importancia del predictor*, el mismo que indica el grado de aportación de cada variable con respecto a la variable *Destino*.

Figura 66: IA (Microcrédito) - Resultado, Gráfico de Modelos generados

El detalle de las características de cada modelo se encuentra en el **Anexo *5*:** Características del modelo diseñado con IA para microcréditos.

5.2.5 Validación del modelo

Tabla de comparación del valor observado con el pronosticado

En la tabla comparación de $XF-riesgoPago (Pronosticado) con riesgoPago (Observado) de la Figura 67, se muestra que los resultados obtenidos en cada una de las partes son alentadores, aproximadamente en promedio el 92% de resultados son correctos, esto se refiere a la precisión del modelo. Pero lo más importante en esta tabla es que, estos resultados se conservan para cada una de las etapas de la construcción del modelo, lo cual genera confianza en el modelo. En otras palabras, significa que el modelo va a predecir con el mismo nivel de precisión a partir de nuevos casos.

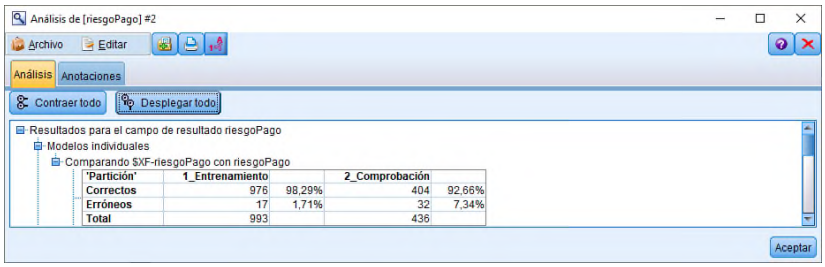


Figura 67: IA (Microcrédito) - Matriz de confusión

Matriz de coincidencias

Además de lo expuesto anteriormente en la matriz de comparación, se observa en la Figura 68 con mayor detalle la clasificación para buenos y malos; para este caso se reafirma lo expresado anteriormente. Tal como hemos definido anteriormente, los valores de "NO" son para créditos malos y "SI" para créditos buenos. El modelo tiene mayor acierto en los buenos (94,87%), para los malos (88,23%). Lo más importante es que en la comprobación los porcentajes de aciertos se mantiene siendo superior los aciertos de buenos sobre malos, a su vez existe una mínima variación por la proporción de registros.

Matriz de coincidencias para $XF-riesgoPago (las filas muestran las reales)

'Partición' = 1_Entrenamiento	NO	SI
NO	423	9
SI	8	553
'Partición' = 2_Comprobación	**NO**	**SI**
NO	170	20
SI	12	234

Figura 68: IA (Microcrédito) - Matriz de coincidencias

Evaluación del rendimiento

En la Figura 69, en comparación con modelos tradicionales, el rendimiento en la categoría malos es muy superior.

Evaluación del rendimiento

'Partición' = 1_Entrenamiento	
NO	0,814
SI	0,555
'Partición' = 2_Comprobación	
NO	0,762
SI	0,49

Figura 69: IA (Microcrédito) - Evaluación del rendimiento

Métricas de evaluación

Los valores de las métricas de evaluación se muestran en la Figura 70.

Métricas de evaluación

'Partición'	1_Entrenamiento		2_Comprobación	
Modelo	AUC	Gini	AUC	Gini
$XF-riesgoPago	0,999	0,997	0,971	0,943

Figura 70: IA (Microcrédito) - Métricas de evaluación

Evaluación SFX RiesgoPago

Los diagramas de evaluación mostrados en las Figura 71 y Figura 72, muestran la calidad con que los modelos predicen resultados particulares.

Tal como se puede observar en ambos gráficos, el área bajo la curva es similar tanto para el entrenamiento, como para la comprobación y el área cubierta es bastante cercana a la ideal.

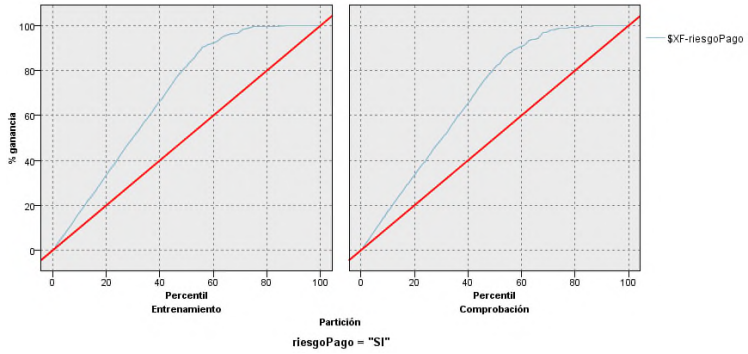


Figura 71: IA (Microcrédito) - Resultado, Evaluación y comparación, Créditos buenos

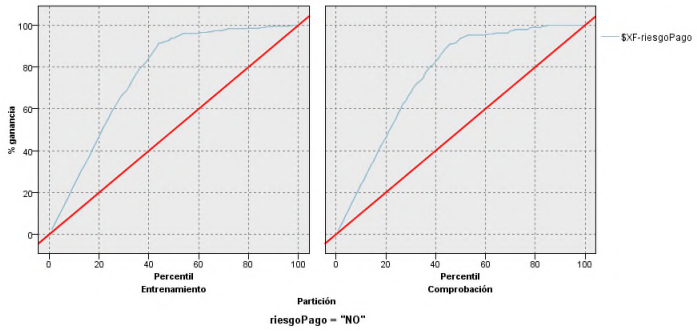


Figura 72: IA (Microcrédito) - Resultado, Evaluación y comparación, Créditos malos

5.3 Comparación de los modelos

En la *Tabla 8: Principales características de los modelos,* se muestran las particularidades de los modelos más relevantes para nuestro trabajo: % de presión general del modelo, que es una métrica dada por la herramienta usada para el desarrollo de los modelos; uso en producción, siendo este individual o integrados en un conjunto llamado Nugget; y las variables que más influencia tuvieron en la caracterización del modelo.

Tabla 8: Principales características de los modelos

Tipo	Tecnicas utilizadas	Precisión general (%)	Uso en Producción	Variables importantes en conjunto
Consumo	**Tradicionales**			
	Regresion Logística	86,15	Individual	Plazo operación
	Inteligencia Artificial			
	C5.1	82,617	Integrados en un Nugget	Plazo operación
	Árbol C&R	82,227		Marca y Clase Vehiculo
	CHAID	81,445		% Entrada, Monto
Microcrédito	**Inteligencia Artificial**			
	Bosque aleatorio	86,927	Integrados en un Nugget	Todas casi por igual
	Lista de decisiones	77,294		
	Árbol XGBoost	92,89		

Para la comparación de los modelos a continuación presentaremos un resumen de las tablas mostradas individualmente en el desarrollo de cada modelo.

Tabla de comparación del valor observado con el pronosticado

En la *Tabla 9: Comparación del valor observado con el pronosticado de los modelos*Tabla 9: Comparación del valor observado con el pronosticado de los modelos, tanto para la etapa de entrenamiento como para la de comprobación, se muestran los resultados obtenidos para cada modelo.

Tabla 9: Comparación del valor observado con el pronosticado de los modelos

Tipo	Tecnicas utilizadas	Tabla de comparación				
			Entrenamiento		Comprobación	
Consumo	Tradicionales	Evaluación	Casos	%	Casos	%
		Correctos	1.014	86,15	419	79,66
		Erroneos	163	13,85	107	20,34
		Totales	**1.177**		**526**	
	Inteligencia Artificial	Evaluación	Casos	%	Casos	%
		Correctos	960	83,99	429	83,95
		Erroneos	183	16,01	82	16,05
		Totales	**1.143**		**511**	
Microcrédito	Inteligencia Artificial	Evaluación	Casos	%	Casos	%
		Correctos	976	98,29	404	92,66
		Erroneos	17	1,71	32	7,34
		Totales	**993**		**436**	

Podemos observar que para los modelos de Inteligencia Artificial los resultados son mejores, pues se mantienen casi los mismos valores en cada una de las etapas de la construcción del modelo, lo cual genera confianza en el modelo. En otras palabras,

significa que estos modelos van a predecir con el mismo nivel de precisión los nuevos casos.

Matrices de coincidencias

En el modelo de Inteligencia Artificial de microcrédito, tal como se ve en la *Tabla 10: Matrices de coincidencias de los modelos,* además de tener un mejor acierto en la determinación de buenos y malos créditos, los porcentajes de aciertos se mantienen similares para cada una de estas categorías, lo que implica que este tipo de modelos bien pueden servir para generar campañas de mercadeo que fomenten la inclusión financiera usando las características encontradas para la determinación de los buenos créditos,

Tabla 10: Matrices de coincidencias de los modelos

Tipo	Tecnicas utilizadas	Matrices de coincidencias						
			Entrenamiento			Comprobación		
Consumo	Tradicionales	Creditos	Malos	Buenos	% acierto	Malos	Buenos	% acierto
		Malos	491	73	87%	198	48	80%
		Buenos	90	523	85%	46	21	31%
	Inteligencia Artificial	Creditos	Malos	Buenos	% acierto	Malos	Buenos	% acierto
		Malos	512	53	91%	227	25	90%
		Buenos	130	448	78%	57	202	78%
Microcrédito	Inteligencia Artificial	Creditos	Malos	Buenos	% acierto	Malos	Buenos	% acierto
		Malos	423	9	98%	170	20	89%
		Buenos	8	553	99%	12	234	95%

Evaluación del rendimiento

En la Tabla 11: Evaluación del rendimiento de los modelos, el modelo de Inteligencia Artificial de microcrédito en comparación con modelos tradicionales, el rendimiento en la categoría malos es muy superior.

Tabla 11: Evaluación del rendimiento de los modelos

Tipo	Tecnicas utilizadas	Evaluación del rendimiento		
Consumo	Tradicionales	Creditos	Entrenamiento	Comprobación
		Malos	0,567	0,523
		Buenos	0,522	0,459
	Inteligencia Artificial	Creditos	Entrenamiento	Comprobación
		Malos	0,478	0,483
		Buenos	0,57	0,563
Microcrédito	Inteligencia Artificial	Creditos	Entrenamiento	Comprobación
		Malos	0,814	0,762
		Buenos	0,555	0,49

Métricas de evaluación

Los valores de las métricas que se muestran en la

Tabla 12: Métricas de evaluación de los modelos demuestran que el modelo de Inteligencia Artificial para microcrédito es mucho mejor tanto para la determinación de los buenos y malos créditos.

Tabla 12: Métricas de evaluación de los modelos

Metricas de evaluación		
Métrica	Entrenamiento	Comprobación
AUC	0,942	0,866
Gini	0,884	0,733
Métrica	Entrenamiento	Comprobación
AUC	0,903	0,909
Gini	0,805	0,819
Métrica	Entrenamiento	Comprobación
AUC	0,999	0,971
Gini	0,997	0,943

5.4 Implementación del modelo

Este es el último paso del desarrollo del scoring de crédito y consiste en calcular la puntuación del potencial del cliente, considerando para esto las variables seleccionadas y el grado y signo asignado a cada variable.

De esta manera se realiza el análisis de crédito de un solicitante, se verifica la puntuación obtenida dentro de los rangos o políticas de crédito y se decide otorgar o negar el crédito.

Las ventajas de la implementación de modelos de score de crédito se las puede resumir tal como en la Tabla 13**.**

Tabla 13: Ventajas y Desventajas del Modelo de Score de Crédito

VENTAJAS	DESVENTAJAS
Calcula la probabilidad de incumplimiento de un crédito; por tanto el nivel de riesgo que debe asumir una institución financiera.	Requiere de una extensa sólida e histórica base de datos. Con características que influyan en el comportamiento de pago del cliente.
Clasifica a un cliente entre buen o mal sujeto de crédito y así permite tomar la decisión de aprobar o rechazar la solicitud de crédito.	Se debe contar con personal que tenga conocimiento y experiencia en el desarrollo e implementación del modelo para el monitoreo, caso contrario la dependencia de un consultor externo trae consigo un riesgo operativo.
Es una forma transparente de calificar al solicitante, con base a sus características y capacidad de pago.	
Determina las posibles políticas de crédito que puede aplicar una institución financiera.	Depende de la implementación de un sistema transaccional y del personal asignado al proyecto con no debe cometer errores en el proceso; caso contrario el modelo puede fallar en esta etapa.
Disminuye el tiempo de análisis de una solicitud de crédito y aprobación del crédito.	
El modelo puede ser validado en cualquier momento y es seleccionado de acuerdo a los factores estadísticos que mejor se ajusten al caso de análisis.	La base de datos no contiene información de cartera rechazada y está contendrá únicamente datos cartera aceptada de acuerdo a políticas anteriores.
Mejora el tipo de cartera que tiene una institución financiera y asigna los recursos de mejor manera.	El modelo no considera o pronostica casos que no hayan sucedido con alta frecuencia en el pasado y sólo destaca casos de alto riesgo.
Permite estimar pérdidas o rentabilidad sobre las colocaciones que realice la institución.	Excluye los casos cuyos campos están en blanco en alguna variable.
Orienta a la institución a mercados más rentables y menos riesgosos.	El pasado no siempre es un buen estimador del futuro, la base de datos sobre la cual se desarrolla el modelo contiene información del pasado y sobre cartera aprobada con políticas anteriores.
Disminuye el tiempo y costo de cobranzas al seleccionar cartera con menor probabilidad de incumplimiento.	

Fuente: Simbaña M. Sandra (2012).

6. Sistema de información

A partir de la etapa de preparación de los datos, en la metodología CRISP-DM, con los conjuntos de datos obtenidos se puede realizar gráficos que adecuadamente organizados conforman un Sistema de Información (*dashboard)* para visualización.

Si regresamos hacia la preparación de datos, recordaremos que uno de los objetivos principales de esta atapa es la obtención de nuevas variables que puedan aportar significativamente al entendimiento del comportamiento, si luego incorporamos el conocimiento de estas variables a los mismos gráficos de un típico dashboard de inteligencia de negocios (BI), estamos concentrando analítica avanzada a este tablero de control, es decir estamos pasando del Business Intellicence a Business Analytics.

Ejemplo de ello es la primera parte del Sistema de Información desarrollado para este proyecto en *PowerBI* a partir de los datos obtenidos en esta, y tal como lo muestra la Figura 73, se trata de gráficos típicos pero que incorporan dentro del mismo la historia de las variables que apoyan en entendimiento del comportamiento:

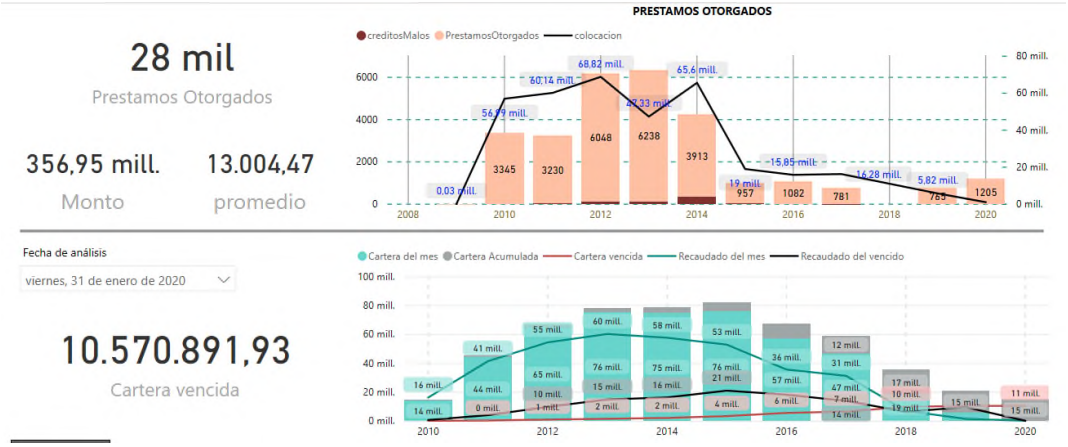


Figura 73:Página Inicial del Sistema de Información

En el primer gráfico podemos ver, en cada una de las barras, información adicional respecto al histórico del número de créditos vencidos y de igual manera para el segundo gráfico información respecto a la recuperación de cartera tanto vencida como del mes.

Como lo vimos en la etapa de modelización, siempre en la selección de las técnicas que integran un modelo Nugget, al menos hay un Árbol de Decisiones, de cual obtenemos información que utilizamos para la segunda parte del Sistema de Información como se detalla a continuación en la Figura 74, donde se visualizan gráficos donde se analiza la calificación dada por el modelo a todas las solicitudes durante el proceso de calificación.

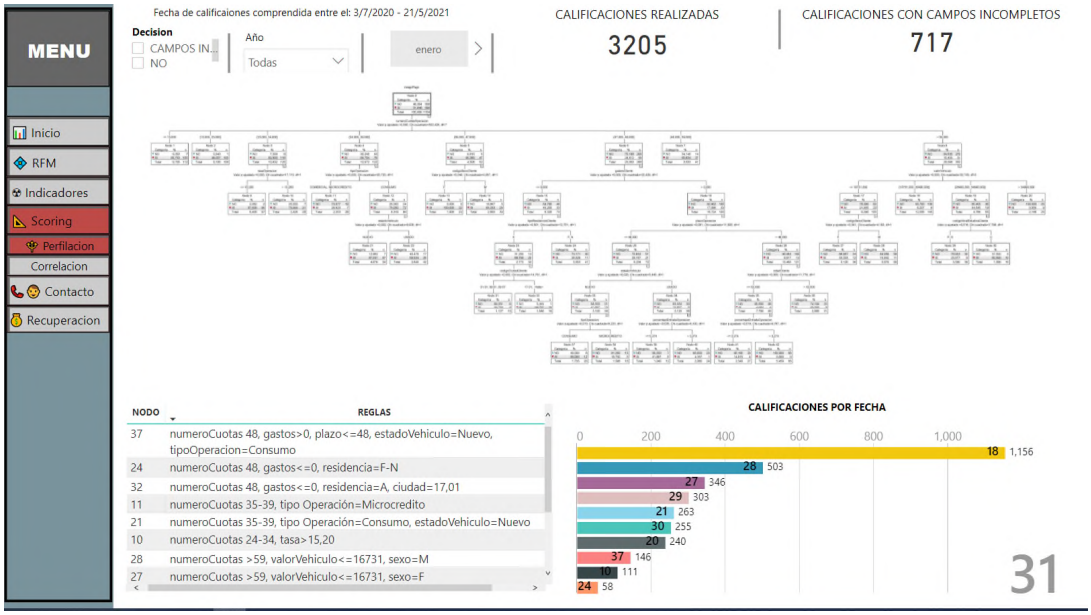


Figura 74: Perfilación de las solicitudes de crédito

Primero se visualiza un gráfico con el esquema del Árbol de Decisiones, donde cada rama o nodo significa un perfil que agrupa a un determinado grupo de prospectos que tienen características similares.

En la parte inferior izquierda se presenta en orden de frecuencias los nodos que más participaron en las puntuaciones, para cada nodo se muestran las reglas de que lo definen.

En la parte inferior derecha vemos un gráfico de barras dinámico, donde cada barra representa a un Nodo final y su altura el número de veces que el perfil de ese Nodo participo en una solicitud, ya sea para un SI o un NO.

El gráfico dinámico muestra el crecimiento o lo contrario de la frecuencia de participación de un perfil en el transcurrir del tiempo de análisis.

Probablemente un descubrimiento será comprobar que gran parte de las solicitudes presentadas no coinciden con el enfoque que se pretendía con las campañas de mercadeo.

Si tomamos muy en cuenta aquellos perfiles donde el porcentaje de las solicitudes históricamente dieron buenos resultados y los contrastamos con el numero proporcional de veces que las solicitudes coincidían con ese perfil, el departamento de mercadeo podrá reorientar sus campañas.

También este análisis podrá ayudar a descubrir nuevos nichos de mercado que inicialmente no fueron tomados en cuanta en las campañas de mercadeo.

Finalmente, en una tercera parte del *dashboard*, incluimos de desarrollo de un Sistema de Recomendación, tal como se muestra en la Figura 75, donde al igual que en la sección anterior se muestra un gráfico con el esquema de un Árbol de Decisiones, pero que para este caso será el de un Árbol Interactivo, es decir que se pueda podar, para que sus ramas hagan énfasis en las variables de interés particular, como en ciertas características del crédito o las del bien a financiar, recomendaciones que normalmente puedan ser tomadas para mejorar la probabilidad de pago en oficio de casos históricos.

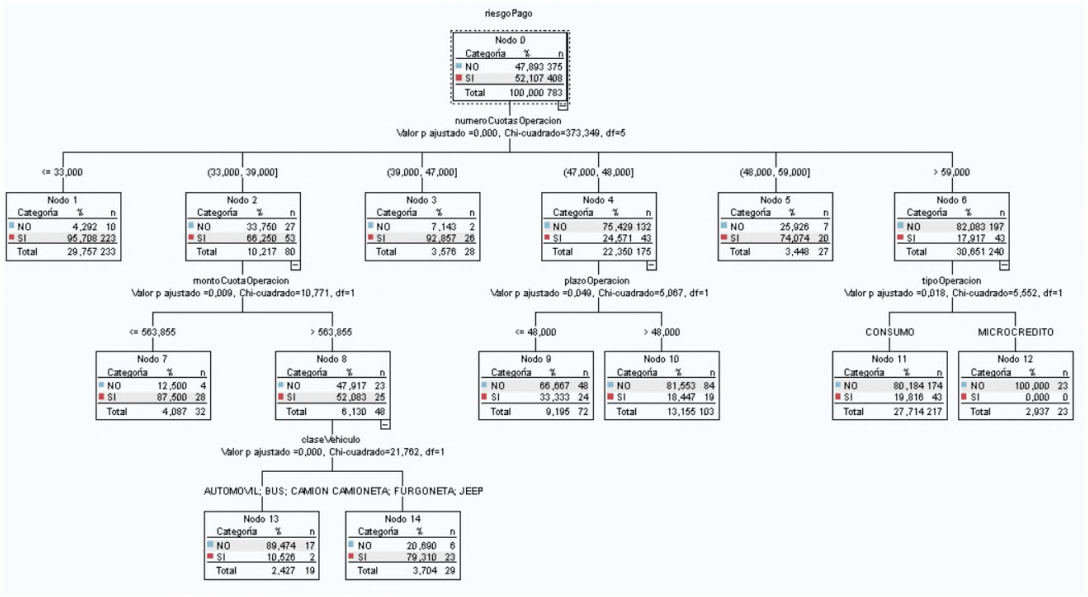


Figura 75: Árbol de Decisiones Interactivo

De este Árbol Interactivo se obtiene las ganancias históricas de los nodos que tuvieron un resultado positivo, ordenados en deciles según la población de cada uno de ellos, tal como se ve en la **Tabla 14**.

Tabla 14: Ganancias históricas de los nodos

Conjunto de desarrollo de árboles

Nodos	Percentil	Percentil: n	Ganancia: n	Ganancia (%)	Respuesta (%)	Índice (%)
1	10,00	78,00	75,00	18,30	95,71	183,68
1	20,00	157,00	150,00	36,83	95,71	183,68
1,3	30,00	235,00	225,00	55,11	95,68	183,63
3,7,14	40,00	313,00	293,00	71,78	93,57	179,56
14,5,9	50,00	392,00	334,00	81,94	85,29	163,68
9,11	60,00	470,00	354,00	86,69	75,26	144,43
11	70,00	548,00	369,00	90,48	67,37	129,28
11	80,00	626,00	385,00	94,27	61,44	117,91
11,10	90,00	705,00	399,00	97,88	56,65	108,71
10,13,12	100,00	783,00	408,00	100,00	52,11	100,00

Cuando el modelo está en producción y determina la viabilidad de una solicitud pero que el Nodo que hace referencia esta calificación no está en una posición suficientemente arriba de esta tabla, este sistema recomienda cuales son las modificaciones que se podrían hacer, que obviamente sean viables, para ascender la respuesta a nodos superiores, aumentando así la Ganancia (%) y el Índice (%) en la solicitud.

7. Conclusiones y recomendaciones

CONCLUSIONES

1. Si bien en un crédito de consumo es fundamental observar la capacidad de pago, un excesivo enfoque en esta variable no aclara la realidad de que en muchos casos la falta de pago tiene relación con el destino y características del bien financiado.

2. Por más que se trate de un crédito de consumo, el enfoque en la capacidad de pago no fomenta la inclusión social al sistema financiero. Para paliar este efecto negativo, echamos mano a las técnicas de modelamiento basadas en Inteligencia Artificial, que encuentran un mayor aporte en las variables no relacionadas con la capacidad financiera de los prospectos. En este proyecto, se han utilizado variables relacionadas con el objeto a financiar, que no son datos de carácter personal y por lo tanto no existe la posibilidad de atentar contra la privacidad de los solicitantes de un crédito

3. La masificación de los datos hace que para su tratamiento se requieran de técnicas que aprovechen de mejor manera las capacidades computacionales actuales y esto es posible conseguirlo, mediante el uso de técnicas de Inteligencia Artificial.

4. En el caso de los microcréditos, la evaluación previa a la aprobación, se la debería realizar en base a la actividad económica del proyecto para el que se solicita el crédito, pero en la práctica el proceso de aprobación se lo maneja como si se tratase de un crédito de consumo, esto es, tomando en cuenta únicamente la capacidad de pago de quien lo solicita. El presente trabajo deja claro que es posible determinar si es factible o no el otorgamiento de microcréditos, utilizando variables relacionadas con la actividad económica del proyecto para el que se lo solicita.

5. La adquisición de un vehículo podría estar orientada para una actividad comercial o para uso familiar, incluso ambos, y esto es algo que el modelo debe considerar.

6. En este proyecto se diseñó un artefacto que mejora el proceso de calificación para el otorgamiento de un crédito automotriz, para lo cual primero plantea una metodología técnica para el tratamiento de los datos y una metodología para la adquisición y cuantificación de la información de los prospectos en el proceso de calificación. Así mismo, la evaluación de este método enfatiza el destino y características del bien que será financiado.

7. El modelo, basado en técnicas de Inteligencia Artificial para microcrédito que finalmente se obtuvo en el presente trabajo, se comporta de forma muy adecuada para los créditos automotrices, además el análisis de sus resultados puede ser usado para generar campañas de mercadeo que fomenten la inclusión financiera haciendo uso de las características encontradas para la determinación de los buenos créditos.

8. Las variables que representan las características del bien a financiar aportan significativamente al modelo matemático, demostrando su influencia en la determinación del riesgo de pago de un crédito automotriz.

9. Finalmente, al menos una de las tres técnicas que integran en conjunto el modelo, es un Árbol de Decisión, cuyos resultados pueden ser visualizados y analizados en un Sistema de Información.

RECOMENDACIONES

1. Se recomienda que a medida que se vayan receptando nuevas solicitudes, toda la información que se pueda recabar de ellas sea almacenada independientemente de si el crédito es otorgado o no, y para el caso en que sean otorgados, estos sean monitoreados permanentemente por un modelo de comportamiento en la cobranza, de tal manera que se pueda desde un inicio evaluar la eficiencia de la metodología y del modelo.

2. Durante el proceso de recolección y evaluación para la calificación de crédito, se podrían incluir variables que no estuvieron presentes al momento de construcción del modelo, por lo tanto, se recomienda que las mismas sean incorporadas en el reentrenamiento de los modelos.

3. Se recomienda la construcción de varios modelos, de tal manera que los modelos sean más específicos y personalizados para cada ámbito de negocio.

4. Finalmente, se debe acompañar el proceso de tratamiento de los datos con un Sistema de Información adecuado a las necesidades de la institución financiera.

8. Bibliografía

[1] T. W. B. Group, «Credit Scoring Approaches Guidelines», *Encycl. Financ.*, pp. 76-76, 2019, doi: 10.1007/0-387-26336-5_521.

[2] S. de Bancos, «Ley Organica de Instituciones del Sistema Financiero», Quito, 2014. Accedido: feb. 16, 2020. [En línea]. Disponible en: www.lexis.com.ec.

[3] A. Garcia Serrano, Inteligencia Artificial. Fundamentos Prácticas y Aplicaciones, Segunda. Mexico: Alfaomega, 2016.

[4] A. Fernandez, «Artificial Intelligence in Financial Services», *SSRN Electron. J.*, n.º March, 2019, doi: 10.2139/ssrn.3366846.

[5] M. M. Chlebus, «Towards better understanding of complex machine learning models using explainable artificial intelligence (xai) -case of credit scoring modelling», n.º July, 2020, doi: 10.13140/RG.2.2.25170.79041.

[6] S. Guerrón Ayala, «Análisis y preparación estadística de variables para el diseño de un modelo credit score de gestión de riesgo de crédito», Universidad Andina Simón Bolívar Sede Ecuador, 2008.

[7] S. Simbaña, «Desarrollo de un Score de crédito para el financiamiento automotríz, con base en el análisis estadístico de variables», Universidad Andina Simón Bolívar, 2012.

[8] S. Muñoz, «Analisis del Credit Scoring», p. 102, 2012, [En línea]. Disponible en: http://repositorio.usfq.edu.ec/bitstream/23000/2185/1/106073.pdf.

[9] F. Cordova, «La gestión de riesgos en las cooperativas de ahorro y crédito», *J. Coll. Stud. Dev.*, vol. 1, n.º 6, p. 9, 2010.

[10] C. Iñiguez y M. Morales, «Selección de perfiles de clientes medientae Regresión Logística para muestras desproporcionadas, validación, monitoreo y aplicación en la proyección de provisiones», Escuela Politécnica Nacional, 2009.

[11] B. Baesens, M. Egmont-Petersen, R. Castelo, y J. Vanthienen, «Learning Bayesian network classifiers for credit scoring using Markov Chain Monte Carlo search», *Proc. - Int. Conf. Pattern Recognit.*, vol. 16, n.º 3, pp. 49-52, 2002, doi: 10.1109/icpr.2002.1047792.

[12] F. Of, «The future of credit scoring modelling using advanced techniques», *Sci. Am.*, n.º February, pp. 38-59, 2020.

[13] M. Bensic, N. Sarlija, y M. Zekic-Susac, «Modelling small-business credit scoring by using logistic regression, neural networks and decision trees», *Intell. Syst. Accounting, Financ. Manag.*, vol. 13, n.º 3, pp. 133-150, 2005, doi: 10.1002/isaf.261.

[14] D. Kumar Gupta y S. Goyal, «Credit Risk Prediction Using Artificial Neural Network Algorithm», *Int. J. Mod. Educ. Comput. Sci.*, vol. 10, n.º 5, pp. 9-16, 2018, doi: 10.5815/ijmecs.2018.05.02.

[15] M. M. Amaro, «Credit Scoring : Comparison of Non-Parametric Techniques against Logistic Regression», 2020.

[16] J. B. Simha, «Comparing decision trees with logistic regression for credit risk analysis», *Int. Inst. Inf. Technol. Bangalore India*, vol. 7, n.º January 2006, 2006, [En línea]. Disponible en: http://www.abibasystems.com/white_paper/credit_risk.pdf.

[17] J. Bastos, «Credit scoring with boosted decision trees Credit scoring with boosted decision trees», n.º 8034, 2008.

[18] H. Ince y B. Aktan, «A comparison of data mining techniques for credit scoring in banking: A managerial perspective», *J. Bus. Econ. Manag.*, vol. 10, n.º 3, pp. 233-240, 2010, doi: 10.3846/1611-1699.2009.10.233-240.

[19] D. Zhang, X. Zhou, S. C. H. Leung, y J. Zheng, «Vertical bagging decision trees model for credit scoring», *Expert Syst. Appl.*, vol. 37, n.º 12, pp. 7838-7843, 2010, doi: 10.1016/j.eswa.2010.04.054.

[20] G. Wang, J. Ma, L. Huang, y K. Xu, «Two credit scoring models based on dual strategy ensemble trees», *Knowledge-Based Syst.*, vol. 26, pp. 61-68, 2012, doi: 10.1016/j.knosys.2011.06.020.

[21] S. E. E. Profile, «A New Credit Scoring Method Based on Rough Sets and Decision Tree», *Adv. Knowl. Discov. Data Mining, Pacific-Asia Conf. Proc.*, vol. 5012, n.º May 2008, pp. 0-4, 2014, doi: 10.1007/978-3-540-68125-0.

[22] F. Shahbazi, «Using decision tree classification algorithm to design and construct the credit rating model for banking customers», vol. 21, n.º 3, pp. 24-28, 2019, doi: 10.9790/487X-2103022428.

[23] F. Innocenti, «Machine Learning in Credit Scoring», April, 2020.

[24] P. Golbayani, I. Florescu, y R. Chatterjee, «A comparative study of forecasting Corporate Credit Ratings using Neural Networks, Support Vector Machines, and Decision Trees», 2020, [En línea]. Disponible en: http://arxiv.org/abs/2007.06617.

[25] C. L. Huang, M. C. Chen, y C. J. Wang, «Credit scoring with a data mining approach based on support vector machines», *Expert Syst. Appl.*, vol. 33, n.º 4, pp. 847-856, 2007, doi: 10.1016/j.eswa.2006.07.007.

[26] A. Ghodselahi y A. Amirmadhi, «Application of Artificial Intelligence Techniques for Credit Risk Evaluation», *Int. J. Model. Optim.*, vol. 1, n.º 3, pp. 243-249, 2011, doi: 10.7763/ijmo.2011.v1.43.

[27] A. Chacko, A. Antonidoss, y A. Sebastain, «Optimized algorithm for credit scoring», *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, n.º 1.3 Special Issue, pp. 361-365, 2020, doi: 10.30534/ijatcse/2020/5691.32020.

[28] S. F. Crone y S. Finlay, «Instance sampling in credit scoring: An empirical study of sample size and balancing», *Int. J. Forecast.*, vol. 28, n.º 1, pp. 224-238, 2012, doi: 10.1016/j.ijforecast.2011.07.006.

[29] N. Mohammadi y M. Zangeneh, «Customer Credit Risk Assessment using Artificial Neural Networks», *Int. J. Inf. Technol. Comput. Sci.*, vol. 8, n.º 3, pp. 58-66, 2016, doi: 10.5815/ijitcs.2016.03.07.

[30] S. Imtiaz y A. J., «A Better Comparison Summary of Credit Scoring Classification», *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, n.º 7, 2017, doi: 10.14569/ijacsa.2017.080701.

[31] A. Natasha, D. D. Prastyo, y Suhartono, «Credit scoring to classify consumer loan using machine learning», *AIP Conf. Proc.*, vol. 2194, n.º December, 2019, doi: 10.1063/1.5139802.

[32] L. Gambacorta, Y. Huang, y H. Qiu, «How do machine learning and non-traditional data affect credit scoring?», n.º 834, 2019.

[33] N. R. Tadapaneni, «Artificial Intelligence in Finance», *Int. J. Innov. Res. Sci. Eng. Technol.*, vol. 9, n.º 5, pp. 1-29, 2020, doi: 10.5281/zenodo.2612537.

[34] C. Edmond, I. Journal, C. Edmond, y A. S. Girsang, «Classification Performance for Credit Scoring using Neural Network», *Int. J. Emerg. Trends Eng. Res.*, vol. 8, n.º 5, 2020.

[35] A. Cao, H. He, Z. Chen, y W. Zhang, «Performance Evaluation of Machine Learning Approaches for Credit Scoring», *Int. J. Econ. Financ. Manag. Sci.*, vol. 6, n.º 6, pp. 255-260, 2018, doi: 10.11648/j.ijefm.20180606.12.

[36] D. Guégan y B. Hassani, «Regulatory learning: How to supervise machine learning models? An application to credit scoring», *J. Financ. Data Sci.*, vol. 4, n.º 3, pp. 157-171, 2018, doi: 10.1016/j.jfds.2018.04.001.

[37] B. Niu, J. Ren, y X. Li, «Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending», *Inf.*, vol. 10, n.º 12, 2019, doi: 10.3390/INFO10120397.

[38] B. W. Yap, S. H. Ong, y N. H. M. Husain, «Using data mining to improve assessment of credit worthiness via credit scoring models», *Expert Syst. Appl.*, vol. 38, n.º 10, pp. 13274-13283, 2011, doi: 10.1016/j.eswa.2011.04.147.

[39] K. Kennedy, «Credit Scoring Using Machine Learning», 2013, doi: 10.21427/D7NC7J.This.

[40] D. Wu, D. L. Olson, y A. Dolgui, «Artificial intelligence in engineering risk analytics», *Eng. Appl. Artif. Intell.*, vol. 65, pp. 433-435, 2017, doi: 10.1016/j.engappai.2017.09.001.

[41] R. Saia, S. Carta, D. R. Recupero, G. Fenu, y M. Saia, «A discretized enriched technique to enhance machine learning performance in credit scoring», *IC3K 2019 - Proc. 11th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag.*, vol. 1, n.º September, pp. 202-213, 2019, doi: 10.5220/0008377702020213.

[42] B. Sabeti, H. A. Firouzjaee, S. Safavi, y W. Wang, «Credit Risk Rating Using State Machines and Machine Learning», n.º Icefr, 2020.

[43] M. R. Kumar, «Machine Learning-based Credit Scoring System and Framework of " Peer Trust Score " Model 1 Introduction 2 Related Work», vol. 8, n.º 5, pp. 1684-1692, 2019.

[44] D. Zhang, H. Huang, Q. Chen, y Y. Jiang, «A comparison study of credit scoring models», *Proc. - Third Int. Conf. Nat. Comput. ICNC 2007*, vol. 1, n.º Icnc, pp. 15-18, 2007, doi: 10.1109/ICNC.2007.15.

[45] G. Wang, J. Hao, J. Ma, y H. Jiang, «A comparative assessment of ensemble learning for credit scoring», *Expert Syst. Appl.*, vol. 38, n.º 1, pp. 223-230, 2011, doi: 10.1016/j.eswa.2010.06.048.

[46] M. Khanbabaei y M. Alborzi, «The Use of Genetic Algorithm, Clustering and Feature Selection Techniques in Construction of Decision Tree Models for Credit Scoring», *Int. J. Manag. Inf. Technol.*, vol. 5, n.º 4, pp. 13-32, 2013, doi: 10.5121/ijmit.2013.5402.

[47] F. Louzada, A. Ara, y G. B. Fernandes, «Classification methods applied to credit scoring: Systematic review and overall comparison», *Surv. Oper. Res. Manag. Sci.*, vol. 21, n.º 2, pp. 117-134, 2016, doi: 10.1016/j.sorms.2016.10.001.

[48] Y. L. Eddy, E. Muhammad, N. Engku, y A. Bakar, «Credit scoring models: Techniques and issues», vol. 2, n.º 2, pp. 29-41, 2017.

[49] S. Walusala, R. Rimiru, y C. Otieno, «A Hybrid Machine Learning Approach for Credit Scoring Using PCA and Logistic Regression», *Int. J. Comput. Int. J. Comput. (IJC*, vol. 27, n.º 1, pp. 84-102, 2017, [En línea]. Disponible en: http://ijcjournal.org/.

[50] A. Chopra y P. Bhilare, «Application of Ensemble Models in Credit Scoring Models», *Bus. Perspect. Res.*, vol. 6, n.º 2, pp. 129-141, 2018, doi: 10.1177/2278533718765531.

[51] I. Spss, «Guía de aplicaciones de IBM SPSS Modeler 18.1». IBM, p. 368, 2015, doi: 10.1017/CBO9781107415324.004.

[52] G. Fahner, «Developing Transparent Credit Risk Scorecards More Effectively : An Explainable Artificial Intelligence Approach», n.º c, pp. 7-14, 2018.

[53] E.-I. Dumitrescu, S. Hué, C. Hurlin, y sessi tokpavi, «Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds», *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3553781.

[54] I. Spss, «IBM SPSS Collaboration and Deployment Services Deployment Manager 4 . 2 User ' s Guide». p. 368, 2018.

[55] A. R. Hevner, S. T. March, J. Park, y S. Ram, «Design science in information systems research», *MIS Q. Manag. Inf. Syst.*, vol. 28, n.º 1, pp. 75-105, 2004, doi: 10.2307/25148625.

[56] E. Politou, E. Alepis, y C. Patsakis, «Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions», *J. Cybersecurity*, vol. 4, n.º 1, pp. 1-20, 2018, doi: 10.1093/cybsec/tyy001.

[57] R. Cheripelli y K. R. Sri, «Evaluation of machine Learning Models for Credit Scoring», n.º 2798, pp. 2798-2805, 2020.

[58] N. Aggarwal, «The Norms of Algorithmic Credit Scoring», *SSRN Electron. J.*, n.º April, 2020, doi: 10.2139/ssrn.3569083.

[59] T. Darmawan, A. S. Birawa, E. Eryanto, y T. Mauritsius, «Credit classification using crisp-dm method on bank abc customers», *Int. J. Emerg. Trends Eng. Res.*, vol. 8, n.º 6, pp. 2375-2380, 2020, doi: 10.30534/ijeter/2020/28862020.

9. Glosario de términos

Crédito

Definición 1: Créditos de consumo

Los **créditos de consumo** son préstamos que concede una institución financiera para la adquisición de bienes o servicios. Es decir, recoge los créditos otorgados para compras comunes de los hogares, como la compra de un automóvil, muebles, viajes, cualquier otro gasto extra o imprevisto.

En el Ecuador este tipo de crédito se dividen en dos:

Crédito de Consumo Ordinario, es el otorgado a personas naturales, cuya garantía sea de naturaleza prendaria o fiduciaria, con excepción de los créditos prendarios de joyas. Incluye anticipos de efectivo o consumos con tarjetas de crédito corporativas y de personas naturales, cuyo saldo adeudado sea superior a USD 5 000, excepto en establecimientos médicos y educativos.

Crédito de Consumo Prioritario, es el otorgado a personas naturales para la compra de bienes, servicios o gastos no relacionados con una actividad productiva, comercial y otras compras y gastos no incluidos en el segmento de consumo ordinario, incluidos los créditos prendarios de joyas.

Incorpora los anticipos de efectivo o consumos con tarjetas de crédito corporativas y de personas naturales, cuyo saldo adeudado sea hasta USD 5 000; excepto en los establecimientos educativos. Comprende los consumos efectuados en los establecimientos médicos cuyo saldo adeudado por este concepto sea superior a USD 5.000.

Definición 2: Microcrédito

Es el otorgado a una persona natural o jurídica con un nivel de ventas anuales inferior o igual a USD 100.000,00, o a un grupo de prestatarios con garantía solidaria, destinado a financiar actividades de producción y/o comercialización en pequeña escala, cuya fuente principal de pago la constituye el producto de las ventas o ingresos generados por dichas actividades, verificados adecuadamente por la entidad del Sistema Financiero Nacional.

Microcrédito Minorista: Operaciones otorgadas a solicitantes de crédito cuyo saldo adeudado en microcréditos a la entidad del sistema financiero, sea menor o igual a USD 1.000,00, incluyendo el monto de la operación solicitada.

Microcrédito de Acumulación Simple: Operaciones otorgadas a solicitantes de crédito cuyo saldo adeudado en microcréditos a la entidad del sistema financiero nacional sea

superior a USD 1.000,00 y hasta USD 10.000,00, incluyendo el monto de la operación solicitada.

Microcrédito de Acumulación Ampliada: Operaciones otorgadas a solicitantes de crédito cuyo saldo adeudado en microcréditos a la entidad del sistema financiero nacional sea superior a USD 10.000,00, incluyendo el monto de la operación solicitada.

Sujeto de crédito: A quién el Banco considere apto para otorgar un microcrédito

Financieros

Ingreso neto: El dinero que recibe el individuo una vez realizado todos los descuentos.

Nivel de endeudamiento: Es el riesgo asumido por el Banco mide las obligaciones del individuo dentro y fuera de Banco.

Patrimonio: Conjunto de bienes propios de una persona.

Riesgo: Es la posibilidad de que se produzca un hecho generador de pérdidas que afecten el valor económico de las instituciones.

Capacidad de pago: Liquidez con la que cuenta el cliente con relación al ingreso total menos los gastos.

Computación

Definición 3: Base de datos

Las bases de datos es un lugar, generalmente archivos, donde se almacenan datos de entidades y que se pueden administrar a través de programas informáticos. Las entidades, personas, bienes materiales, sucesos, transacciones, clientes o cualquier objeto que tenga características propias. Las características toman el nombre de campo y son parte de la información acerca de la entidad y en conjunto se las conoce como registros; por tanto, toda la información de una persona es un registro.

Las organizaciones a diario generan bases de datos de clientes, compras, pagos, ventas, procesos y cualquier otra transacción que realicen. Estas bases de datos contienen información valiosa y que bien administradas pueden generar dinero y valor para quien las posee y administre. Para esto se debe eliminar la redundancia de datos; es decir eliminar datos repetidos y se debe evitar el ingreso incorrecto de estos, para no tener datos inconsistentes.

Estos datos se pueden transformar en información valiosa que ayuden a la toma de decisiones, previo de un análisis estadístico de las características o variables, que se almacenan en las bases de datos o a través del desarrollo de modelos estadísticos.

Contenido de las Bases de Datos

Como se indicó anteriormente, las bases de datos contienen registros con diversas variables o características, por tanto, se dice que las variables son características de la población, muestra o suceso que se desea estudiar. Estas variables pueden ser cualitativas o cuantitativas, las primeras describen a un individuo, grupo de personas o sucesos, por sus características, cualidades o atributos y que no pueden ser medidos en números; por ejemplo: el género, la profesión, el tipo de empleo y otros.

En cambio, las segundas son aquellas variables que describen a un individuo o suceso a través de una apreciación numérica y que se puede medir, cuantificar y realizar operaciones aritméticas con ellas.

Sobre los datos

Definición 4: Variables Cualitativas

Son de tipo nominal y de tipo ordinal. Las nominales son características de modalidad que no admiten un criterio de orden, pero que permiten identificar a los individuos u objetos para clasificarlos por grupos. Un ejemplo de estas variables es el estado civil, al cual no se le puede ordenar, pero si clasificar por las modalidades de soltero, casado, viudo, divorciado y otros. Las ordinales son aquellas que describen a un sujeto o suceso con modalidades no numéricas, pero sí admiten un orden. Por ejemplo, los puestos de participación de las instituciones que ofrecen crédito automotriz. La primera es GMAC, la segunda es el Banco Guayaquil y la tercera CFC.

Definición 5: Variables Cuantitativas

Pueden ser discretas o continuas. Las discretas son aquellas que toman valores enteros o aislados; es decir no admiten valores intermedios. Por ejemplo: el número de cargas que tiene el sujeto de crédito o el número de vehículos o propiedades que posee. Por otro lado, las continuas toman valores intermedios entre dos números, por ejemplo: el valor de la cuota de un cliente puede ser de $345,23 y de otro $482,10.

Definición 6: Variables de un modelo

Por lo general un modelo es representado por dos variables, una independiente y otra dependiente. La **variable independiente** es aquella cuyo valor no depende del valor de otra variable y pueden ser cuantitativas o cualitativas o ambas a la vez.

En cambio, las **variables dependientes** son aquellas cuyo valor depende del valor que toma la variable independiente. Pueden ser cualitativas y no únicamente se restringe a un sí o no o categorías dicotómicas. Existen variables de respuesta que son cualitativas por naturaleza; por ejemplo: una familia tiene casa propia o no la tiene.

Definición 7: Modelo matemático

Es un algoritmo matemático que permite predecir el riesgo y la probabilidad de pago de una deuda, con base a la información de crédito del cliente y otros factores que pueden influir en el comportamiento de pago.

El puntaje o score de crédito consiste en cuantificar una serie de factores, de tal manera, que al cliente se le asigne un valor de tres cifras. Entre mayor sea el puntaje que reciba el cliente, menor será la probabilidad de que este caiga en mora y mejor será la posibilidad de pago de la deuda. Por lo contrario, una puntuación baja significa, mayor probabilidad de que el cliente caiga en mora y menos posibilidades de recuperar el dinero prestado. Por tanto, el score de crédito nos indica el nivel de riesgo que representa un cliente para una institución financiera y permite tomar dos decisiones: la primera aprobar o negar un crédito y la segunda establecer una tasa de interés de acuerdo con la puntuación obtenida.

El score y la tasa de crédito están directamente relacionados. Entre mayor sea la puntuación de un cliente, menor será la tasa de interés que obtenga sobre una hipoteca, tarjetas de créditos, crédito automotriz y otros. Esto se debe al nivel de riesgo que representa, la forma en que le mira el prestamista y a las políticas establecidas por cada institución.

Las puntuaciones de crédito se basan en análisis estadísticos de los distintos elementos de crédito como: madurez del crédito, montos adeudos, historial de pago, tipos de crédito, entre otros.

En el mercado existen múltiples modelos de score de crédito desarrollados de acuerdo con las necesidades de cada institución. En Estados Unidos el score de crédito utilizado por el 90 % de las instituciones financieras es FICO.

Las puntuaciones de este score están en un rango de 300 y 850 puntos. Al momento de dar una puntuación al cliente, se consideran cinco categorías, cada uno con un peso asignado y estos son: un 35% para el historial de pagos incluyendo los atrasos, el 30% para los montos adeudados, un 15% al tiempo del historial de crédito, se considera el 10 % para los tipos de crédito utilizados y un 10 % para el nuevo crédito, incluyendo los créditos recientes. Información personal o demográfica como: edad, raza, estado civil, ingreso y empleo no es considerado.

Así como la puntuación de un cliente es diferente al de otro, la puntuación de un mismo cliente también puede variar de score a otro. Todo dependerá de los factores utilizados y analizados en cada modelo. Por esta razón cada institución financiera puede desarrollar su propio score, con base a diversas variables y de acuerdo con los niveles de riesgo que desea asumir.

Modelos con técnicas estadísticas tradicionales

Modelos de Regresiones

El análisis de regresión permite plantear una ecuación para conocer, la relación entre dos o más variables y predecir el valor de la variable dependiente en función de la independiente. El tipo de regresión se considerará por el número de variables independientes que intervengan en el análisis.

Definición 8: Modelos de Regresión Lineal

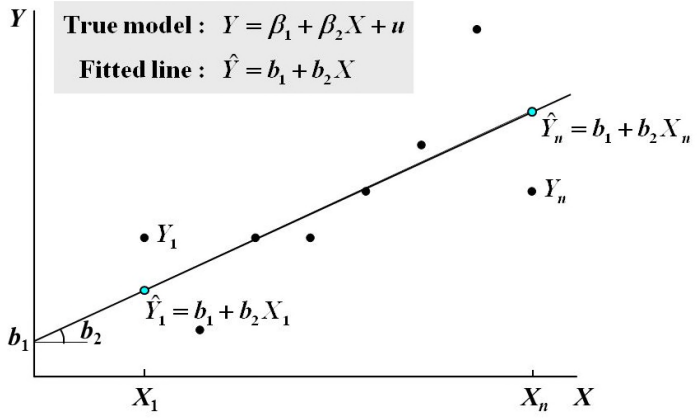
Es un método de análisis de datos que permite conocer la relación entre dos o más variables o pronosticar el comportamiento de una variable a causa de otra o un acontecimiento que afecte a la variable en estudio.

El modelo de regresión lineal simple busca una recta de regresión que relacione a dos variables: una dependiente *Y* y una o varias independientes *X*. Es decir que dicha recta estimará los valores de *Y* que obtendrá para distintos valores de *X*, siendo la fórmula la siguiente:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta_n X_n$$

La regresión lineal busca predecir o estimar el valor promedio de la variable explicada con base a los valores fijos de las variables explicativas. El coeficiente puede tomar valores cercanos a +1 o a -1, para lo cual se elegirá la ecuación de la recta de mínimos cuadrados que mejor se ajuste a la nube de puntos. Estos conceptos son representados en el diagrama de dispersión:

Figura 1: Diagrama de Dispersión



Fuente: Aleta Camejo

Dentro del análisis de la regresión se debe tomar en cuenta, además, el coeficiente de determinación o el *R*2, conocido también como bondad de ajuste. El *R*2 indica el porcentaje de ajuste al usar el modelo lineal; es decir el porcentaje de la variación de *Y*

que se explica a través del comportamiento de *X*; por tanto, a mayor porcentaje mejor es el modelo para predecir el comportamiento de la variable *Y.*

También se puede entender a este coeficiente, como el porcentaje de varianza explicada por la recta de regresión y su valor siempre estará entre 0 y 1 y será igual al cuadrado del coeficiente de correlación ®, entonces:

R2 = r2

Definición 9: Modelo de Regresión lineal simple o de dos variables

Es aquel en que existen únicamente dos variables, una dependiente y otra explicativa o independiente y nos permite conocer la relación que existe entre estas dos variables. El modelo de esta regresión puede ser expresada como:

y = в0 + в1x + s

Donde; *y* es una función lineal de *x*, por tanto, una línea recta.

*β*0*, β*1; son los parámetros del modelo.

s; es una variable aleatoria. "El término de error explica la variabilidad en *y*

que no se puede explicar con la relación lineal entre *x* y".

Definición 10: Modelo de Regresión lineal múltiple

Describe la forma en que la variable dependiente se relaciona con las variables independientes; a través de una ecuación. La regresión lineal múltiple cuenta con varios parámetros, debido a que considerar algunas variables independientes. El modelo general es de la forma:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta_p x_p + s$

Donde; y es una función lineal de x1, x2, ..., x_p.

β0, β1, ..., β_p; son los parámetros de la variable independiente.

s; es una variable aleatoria. "El término de error explica la variabilidad en y

que no se puede explicar el efecto lineal de las p variables independientes".

Definición 11: Regresión no lineal

Son aquellas cuyos parámetros no son lineales, sin importar que las variables sean o no lineales. La linealidad entre dos o más variables es una hipótesis que no se cumple siempre, por lo que existen otro tipo de funciones y modelos para analizar los diferentes casos de estudio.

Definición 12: Modelos de regresión intrínsecamente lineales e intrínsecamente no lineales

Un modelo puede ser lineal en los parámetros y un modelo de regresión lineal en las variables o podemos tener un modelo lineal en los parámetros y no lineal en las variables. Sin embargo, un modelo puede parecer no lineal en los parámetros e intrínseca o inherentemente lineal, debido a que después de una transformación puede convertirse en un modelo de regresión lineal. En el caso de que un modelo no pueda linealizarse en los parámetros, estaremos frente a un modelo intrínsecamente no lineales.

Definición 13: Modelos de regresión de respuesta cualitativa

Se conocen como modelos probabilísticos y consideran que la variable dependiente o regresada puede ser cualitativa, cuantitativa o una mezcla de las dos. Por tanto, la variable de respuesta es una variable binaria o dicótoma porque solo puede tomar dos valores 0 o 1, por ejemplo, será 1 cuando ha estudiado hasta el tercer nivel académico y 0 cuando no ha llegado a ese nivel. Existen cuatro métodos para crear un modelo de probabilidad para una variable de respuesta binaria:

Definición 14: Modelo lineal de probabilidad

Se formula como un modelo de regresión lineal común, pero intrínsecamente se encuentra una probabilidad condicional de que un suceso X_i; tenga lugar, para obtener la variable dependiente Y_i. El modelo se representa en la siguiente función:

$Y_i = \beta 1 + \beta 2 X_i + \mu_i$

Donde:

X; es el ingreso familiar

$Y = 1$; Si la familia tiene casa propia y 0 si no la tiene.

Y_i Dado X_i; $E(Y_i/X_i)$; es decir $Pr(Y_i = 1/X_i)$.

Definición 15: Modelo Logit

Es un modelo de regresión que nos permite conocer la relación entre una o algunas variables independientes cualitativas o cuantitativas y una dependiente o de respuesta binomialmente distribuida y que considera la probabilidad de la ocurrencia de un evento (y) en función de otros factores (x).

Este modelo es similar al lineal el cual intenta explicar la variación o el comportamiento de una variable dependiente; a través de una variable independiente. Sin embargo, en la realidad existen casos de análisis, cuya variable dependiente es afectada por diversos factores que deben ser representados en la ecuación de regresión. La diferencia entre estos dos modelos es la función que utiliza y el uso de las variables ficticias (dummy), que

toman el valor de 0 (malo) y 1 (bueno) y permiten formular una sola ecuación para representar o distinguir diversos grupos de tratamiento.

Este modelo se formula por medio de una ecuación logarítmica que permite clasificar a un individuo de estudio, en un grupo u otro, con base en el análisis de regresión de las variables que afecta o influyen en el comportamiento de los individuos que conforman cada grupo.

La siguiente función define al modelo logit:

$$P\left(Y=\frac{1}{X_i}\right)=\frac{1}{1+e^{-zi}}$$

Donde $Z_i = \beta_0 + \beta_1 X_1 + \mu$

Y las variables se definen de la siguiente forma:

Y_1: Bueno

Y_0: Malo

X_i: Ingreso del cliente

P ($Y = 1/X_i$) = Probabilidad de ser bueno, explicado por la variable X_i.

Z_i: Exponente de una regresión lineal.

β_0 Intercepto de la curva (parámetro a estimar)

β_1: Pendiente de la curva (parámetro a estimar) m: error

i = 1, 2, 3, ...N: índice de diferenciación de variables.

La linealización se realiza utilizando la definición logit, tomando el logaritmo natural de la razón de la probabilidad complementaria, tal como se describe a continuación:

$$L_i = ln\left(\frac{P_i}{1-P_i}\right) = Z_i \qquad ; \qquad L_i = ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 X_i + u$$

Las características del modelo logit son:

- A medida que *P* va de 0 a 1 o *Z* varía de $-\infty$ a $+\infty$, el logit va de $-\infty$ a $+\infty$. Es decir, el logit no está a acotado a estar entre 0 y 1.
- L es lineal en X, pero la probabilidad en sí mismas no lo son.
- Se pueden agregar tantas variables X o regresoras, como lo indique la teoría subyacente.
- Si logit o L es positivo, esto significa que cuando el valor de la regresora se incremente, también se aumente la posibilidad de que las regresadas sean igual a 1. En cambio si L es negativo, las posibilidades de que la regresada sea igual a 1, disminuye conforme el valor de X se incremente.

- La interpretación del modelo dado es: β2, la pendiente mide el cambio de
- L ocasionado por un cambio unitario en X.
- El modelo logit supone que el logaritmo de la razón de probabilidad está relacionado linealmente con Xi.

Definición 16: Modelo Probit

Se basa en la teoría de la utilidad o la perceptiva de selección racional con base en el comportamiento.

Este modelo depende de un índice de conveniencia o variable latente, que se determina por una o algunas variables explicativas. La probabilidad de ocurrencia de un evento (y) está dado por el valor que obtenga el índice; es decir que entre mayor sea el valor del índice, mayor será la probabilidad de ocurrencia.

El modelo probit es de variable dependiente limitada y la estimación de los parámetros se realizan por medio del método de máxima verosimilitud y sugiere tomar los valores de los parámetros que maximicen el logaritmo de la función de verosimilitud.

El modelo se expresa como:

Donde:

$$P(y = 1/x) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n) = G(\beta_0 + \beta X)$$

G, es una función que adopta valores entre cero y uno para los números reales de *z*; además representa la función de distribución acumulativa normal estandarizada por:

$$F(Z_i) = \int_{-\infty}^{\frac{Z_i}{\sigma}} \frac{1}{\sqrt{2\pi}} exp\left[-\frac{t^2}{2}\right] dt$$

Definición 17: Modelo Tobit

Este modelo es una extensión del modelo probit, también se le conoce como modelo de regresión censurada o con variable dependiente limitada. La variable es limitada debido a la restricción impuesta sobre la variable regresada, que necesariamente debe cumplir con una característica, por ejemplo: se desea analizar el nivel de gasto de una familia que ha comprado una casa; por tanto, es necesario que cumpla la condición de tener casa propia para incluirle en el grupo de análisis. El modelo tobit se expresa como:

$$Y_i = \beta_1 + \beta_2 X_i + \mu_i \qquad \text{si LD} > 0$$
$$= 0 \qquad \text{en otro caso}$$

Donde:

LD; lado derecho. Indica que se puede agregar fácilmente otras variables X al modelo.

La regresión con n_1 observaciones y dejar de lado las demás observaciones, no es posible porque n_1 observaciones estarán sesgadas y serán inconsistentes.

Por tanto; "La intuición indica que, si estimamos una regresión basada sólo en las n_1 observaciones, los coeficientes resultantes del intercepto y de la pendiente estarán limitados a ser diferentes de los que obtendríamos si se tomaran en cuenta todas las (n_1 +n_2) observaciones". Para desarrollar este modelo se aplicar el método de máxima verosimilitud.

Validación de los modeles tradicionales

*R*2: Estadístico que expresa el nivel de cambio de la variable dependiente en el modelo. Si el resultado se aproxima a 1, esto significará que tiene un alto nivel de ajuste y capacidad predictiva sobre el caso en estudio. En cambio, los valores cercanos a 0 indican lo contrario.

Los indicadores que se muestran en las tablas de respuesta de la regresión logísticas son:

2 log de la verosimilitud (-2LL): indica hasta qué punto el modelo se ajusta bien a los datos. Entre más pequeño sea el valor mejor será el ajuste.

Valoración de un test diagnóstico en función de sus razones de verosimilitud

Valores de RV Categorización

***RV P* ≥ 10** Incremento **amplio** de la probabilidad de test positivo

5 ≤ *RV P* < 10 Incremento **moderado** de la probabilidad de test positivo

2 ≤ *RV P* < 5 Incremento **pequeño** de la probabilidad de test positivo

1 ≤ *RV P* < 2 Incremento **despreciable** de la probabilidad de test positivo

0,5 < *RV N* ≤ 1 Decremento **despreciable** de la probabilidad de test negativo

0,2 < *RV N* ≤ 0,5 Decremento **pequeño** de la probabilidad de test negativo

0,1 < *RV N* ≤ 0,2 Decremento **moderado** de la probabilidad de test negativo

***RV N* ≤ 0,1** Decremento **amplio** de la probabilidad de test negativo

El R cuadrado de Cox y Snell: es un coeficiente que estima la proporción de la varianza de la variable dependiente explicada por las independientes. Este indicador toma valores entre 0 y 1 y al multiplicarnos por 100 nos indica el porcentaje que tendrá la variable dependiente cuando se incluya una variable al modelo.

El R cuadrado de Nagelkerke: es la versión mejorada del R cuadrado de Cox y Snell. Esta R corrige la escala del estadístico para cubrir el rango de 0 a

1. Este indicador toma un valor inferior máximo de 1

B: Coeficiente de la variable.

S.E: Error estándar de las estimaciones.

Wald: Es un estadístico que aprueba o rechaza la hipótesis nula. Considerando que B = 0. A través de este indicador se comprueba que cada una de las variables sea significativa o no en el modelo. Este indicador debe tomar *p valores* < 0*,* 05 es decir que los parámetros sean distintos de cero; por tanto, significativos.

DF: Son los grados de libertad de cada variable.

Sig: Es el nivel de significancia de Wald. Valor p de significación asociada a cada coeficiente de regresión.

Exp (B): Indica el aumento o disminución de B y permite conocer el nivel de influencia que tiene una variable en el modelo. Este indicador toma valores iguales a 1 cuando la variable no influye en el modelo, menores a uno cuando disminuye la influencia y mayores a uno cuando aumenta la influencia de una variable en el modelo.

Otros métodos de validación pueden ser los de la curva de ROC, Prueba de Kolmogorov – Smirnov o K-S y Prueba Chi que se define a continuación:

Definición 18: Curva ROC

La prueba de Hosmer-Lemeshow ayuda a validar el modelo, calibra e identifica el grado en que la probabilidad predicha coincide con la observada. Además, discrimina y determina el grado en que el modelo distingue la ocurrencia de un evento o no.

Como medida de la discriminación se usa el área bajo la curva ROC. Esta área indica la probabilidad predicha por el modelo y que representa, para todos los pares posibles de individuos formados por un individuo en el que ocurrió el evento y otro en el que no. Por tanto, cuando más alejada está la curva ROC de la diagonal principal mejor es el método de diagnóstico y cuándo más cercana está a la diagonal, el método de diagnóstico es malo.

Al realizar esté análisis en SPSS se obtiene la gráfica y la tabla del área bajo la curva, en esta tabla se detalla el área, cuyo dato debe ser mayor que 0,5; el error estándar debe ser menor a 0.05 con un nivel de confianza del 95 % y el intervalo de confianza se obtendrá al sumar o restar el área de la curva para así tener el límite inferior y superior. En la última tabla encontraremos una lista de las coordenadas de la curva ROC.

Definición 19: Prueba de Kolmogorov – Smirnov o K-S

Es una prueba que nos permite conocer la bondad de ajuste y sirve para confirmar o rechazar la hipótesis nula de que la distribución de una variable se ajusta a una distribución

teórica de probabilidad; es decir que el conjunto de datos sigue una distribución normal. En cambio, la hipótesis alternativa afirma que los datos no siguen una distribución normal. Esta prueba es usada para en muestras superior a 50 registros y se base en evaluar un estadístico:

$$Du = [F_n(x) - F(x)]$$

Dónde:

$Fn(x)$: es la distribución empírica

$F(x)$: es la distribución teórica y que para el caso es la normal.

El valor de este estadístico debe ser menor al error que se aceptará y de acuerdo con un nivel de confianza con el cual se quiere aceptar la hipótesis nula. Cuando esto sucede se aceptará la hipótesis nula y se confirmará que el conjunto de datos sigue una distribución normal.

Para contrastar la hipótesis nula, la prueba K-S se basa en la comparación de dos distribuciones: una empírica y otra teórica. La distribución empírica se obtiene al ordenar de forma ascendente a los datos y se obtiene de la siguiente forma $F(X_i) = i/n$ (donde i es el rango al que corresponde cada observación). Y la distribución teórica depende de la distribución propuesta en la hipótesis.

Definición 20: Prueba Chi-Cuadrado para una muestra

Esta es una prueba de bondad de ajuste que realiza una comparación entre el grupo de frecuencias observadas con el conjunto de frecuencias separadas. Esta comparación nos permite conocer las diferencias entre las dos y evaluar si dos variables están relacionadas o si son independientes. La prueba trabaja con variables categóricas que permiten clasificar los casos por categorías bien definidas y las excluye unas de otras.

La prueba de Chi Cuadrado, indica si la distribución de las frecuencias observadas difiere significativamente de la distribución de las frecuencias que deberíamos esperar; siempre y cuando no hubiese asociación entre dos variables categóricas. La fórmula de cálculo es:

$$X^2 = \frac{\sum (O - E)^2}{E}$$

Dónde:

O es la frecuencia observada.

E es la frecuencia esperada.

La interpretación de este estadístico se basará en el valor de chi-cuadrado, los grados de libertad y su nivel crítico. Cuando el nivel crítico es menor a 0,05 se puede rechazar la

hipótesis de bondad de ajuste y concluir que la variable no se ajusta a una distribución uniforme.

Definición 21: Diseño de la scorecard o tabla de puntajes

Son aquellas tablas que contienen la información sobre las variables seleccionadas, el peso y el signo asignados. Esta información representa al modelo de calificación y que permite identificar el comportamiento de pago. La asignación de peso o valor dado a una variable, nos indica el nivel de importancia que esta tiene en el modelo.

El signo positivo de la variable significa que el atributo premia al cliente cuando este cumple o tiene la característica. En cambio, cuando el signo es negativo y el cliente cumple con este atributo, su calificación disminuirá porque el atributo castiga al cliente.

Definición 22: Determinación de los puntos de corte (cutoff)

Son los rangos dentro de los cuales deben estar las calificaciones de los clientes para la aprobación de un crédito. Estos puntos de corte tienen relación con el grado de exposición o nivel de pérdida que está dispuesta a asumir la institución financiera.

Estos puntos de corte permiten establecer las políticas que crédito y el puntaje mínimo que debe obtener un cliente para que otorgar el crédito. Dependiendo del desplazamiento del rango hacia arriba o hacia abajo, se obtendrá la tasa de rechazo o aprobación en el primer y segundo caso respectivamente.

Definición 23: Red Neuronal Artificial (RNA)

Una Red Neuronal Artificial (RNA) es un modelo simplificado del modo en que los sistemas nerviosos procesan información. Funciona en forma simultánea con un número de unidades simples de procesamiento interconectadas que emulan a las neuronas (llamadas también nodos), están organizadas en niveles denominados capas.

Cada nodo está conectado con otros mediante enlaces de comunicación, cada uno de los cuales tiene asociado un peso. En los pesos se encuentra el "conocimiento" que tiene la RNA acerca de un determinado problema. Algunas de las redes neuronales son herramientas útiles en muchas aplicaciones de predicción en minería de datos debido a su potencia, flexibilidad y facilidad de uso.

Una de las RNA más ampliamente utilizadas en el análisis de clasificación es el Perceptrón Multicapa (MLP por sus siglas en inglés). Rumelhart, Hinton y Williams (1986) formalizaron un método para que una red de este tipo aprendiera la asociación que existe entre un conjunto de patrones de entrada y sus salidas correspondientes. Este método se conoce como backpropagation error (propagación del error hacia atrás).

Un MLP está compuesto por una capa de entrada, una de salida y una o más capas ocultas; aunque se ha demostrado que para la mayoría de los problemas es suficiente con una sola capa oculta Funahashi, 1989.

En este tipo de modelos las conexiones entre nodos siempre van desde las neuronas de una determinada capa hacia las neuronas de la siguiente; no hay conexiones laterales ni hacia atrás. Por tanto, la información siempre se transmite desde la capa de entrada hacia la capa de salida.

Según Palmer, Montaño y Jiménez (2001), en el algoritmo *backpropagation* podemos considerar una etapa donde se presenta ante la red un patrón de entrada y este se transmite a través de las sucesivas capas de neuronas hasta obtener una salida y, por otro lado, una etapa de entrenamiento o aprendizaje en la que se modifican los pesos de la red de manera que coincida la salida deseada con la salida obtenida por la red ante la presentación de un determinado patrón de entrada.

Definición 24: Árboles de Clasificación (Decision Tree)

Los análisis de clasificación basados en árboles de decisión son técnicas de explotación de datos que consisten en estudiar grandes masas de datos con el fin de descubrir patrones no triviales.

Los patrones no triviales que se estudiarán habitualmente serán los predictivos y los explicativos.

Un árbol de decisión representa una serie de pautas basadas en ciertas variables explicativas que se muestran según recorremos el árbol.

Estos árboles se construyen mediante un algoritmo que va dividiendo los registros de la base de datos (casos u observaciones) en nodos de forma recursiva, de manera que con cada subdivisión las frecuencias relativas de las categorías de la variable dependiente vayan tendiendo a 0 o a 1.

IBM SPSS Modeler dispone de seis algoritmos para realizar árboles de clasificación:

- CHAID
- CHAID Exhaustivo
- C&RT o CART
- QUEST
- C5
- Arboles aleatorios

Aunque los árboles de clasificación permiten la construcción de árboles de forma totalmente automatizada, los mejores resultados se obtienen con la colaboración del usuario, al aplicar el conocimiento que tiene de los datos, tomando decisiones racionales al decidir si se va o no a dividir un nodo determinado.

Ventajas de los modelos con IA:

- Transparencia: a diferencia de otros modelos de clasificación, la forma de un árbol es intuitiva y fácil de interpretar.
- Portabilidad: las pautas que se extraen del camino a una hoja del árbol se pueden expresar fácilmente en distintos formatos, como SQL o sintaxis de SPSS.
- Modelización: los modelos de clasificación basados en árboles de clasificación pueden utilizar tanto variables continuas como categóricas; en concreto, si las variables independientes son categóricas y tienen gran número de categorías, entonces estos modelos darán mejores resultados que los modelos de clasificación clásicos.
- No es preciso una habilidad analítica excepcional para “afinar” un árbol de decisión.

Desventajas de los modelos con IA:

En estos modelos se deberá de emplear un gran volumen de datos para asegurarnos que la cantidad de casos en un nodo terminal es significativa.

10. Anexos

Anexo 1: Herramienta utilizada para el análisis de datos

El análisis predictivo reúne capacidades de análisis avanzado que abarcan análisis estadísticos ad hoc, modelado predictivo, extracción de datos, análisis de texto, optimización, puntuación en tiempo real y aprendizaje automático. Estas herramientas ayudan a las organizaciones a descubrir patrones en los datos y van más allá de saber qué ha pasado para anticipar qué es probable que suceda a continuación.

La plataforma IBM Business Analytics, gráficamente ilustrada en la **Figura 76**, ofrece información completa, coherente y precisa en la que confían los encargados de la toma de decisiones para mejorar el rendimiento comercial. Un conjunto integral de inteligencia empresarial, análisis predictivo, rendimiento financiero y gestión de estrategias y aplicaciones de análisis que ofrece una perspectiva clara, inmediata e interactiva del rendimiento actual y la capacidad de predecir resultados futuros. En combinación con extensas soluciones sectoriales, prácticas probadas y servicios profesionales, las organizaciones de cualquier tamaño pueden conseguir el máximo de productividad, automatizar las decisiones de forma fiable y alcanzar mejores resultados.

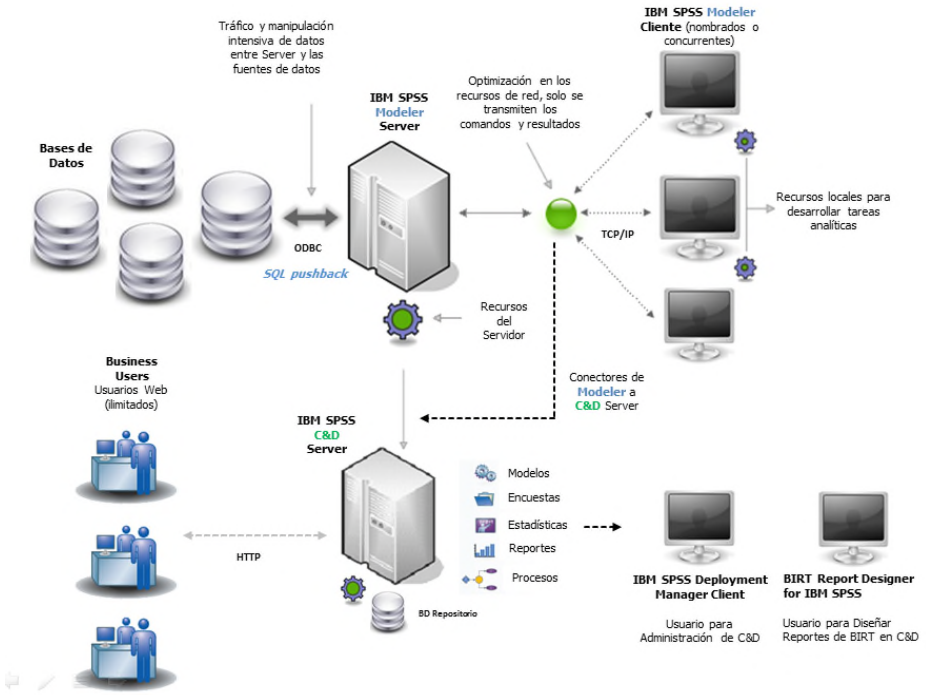


Figura 76: Plataforma para Business Analytics de IBM SPSS

El entorno propuesto tiene capacidades de visualización de los resultados entregados por esta plataforma mediante tableros de control o dashboard, los mismos que podrán

contener filtros de las variables que se defina, para nuestro caso este Sistema de Informacion es desarrollado en Microsoft PowerBI.

IBM® SPSS Modeler Professional (Cliente)

IBM® SPSS Modeler es un conjunto de programas de minería de datos orientado a las empresas.

La interfaz visual de SPSS Modeler permite a los usuarios aplicar su experiencia empresarial específica, lo que deriva en modelos proactivos más eficaces y la reducción del tiempo necesario para encontrar soluciones. SPSS Modeler ofrece muchas técnicas de modelado tales como predicciones, clasificaciones, segmentación y algoritmos de detección de asociaciones.

Tiene una gran variedad de algoritmos para crear modelos fácil e intuitivamente. También trae diversas opciones que permiten visualizar los resultados de tal manera que pueda comunicarlos de forma ágil y sencilla.

Trae múltiples técnicas analíticas dentro de las que se encuentran:

- **Algoritmos de Clasificación**: Elabora predicciones o pronósticos a partir de su información histórica utilizando diversas técnicas como Arboles de Clasificación, Redes Neuronales, Regresión Logística, Series Temporales, Support Vector Machines, regresión de Cox y muchos más.
- **Algoritmos de Segmentación**: Agrupa personas o identifique patrones inusuales con la generación de clúster automáticamente, detección de anomalidades, y opciones de clúster a través de redes neuronales.
- **Algoritmos de Asociación**: Descubre asociaciones, y descubre relaciones o secuencias a través de las diversas técnicas disponibles.

Algoritmos soportados por la herramienta

El nodo Máquina de Vectores de Soporte (SVM), le permite clasificar datos en uno o dos grupos sin que haya un ajuste por exceso. SVM funciona bien con conjuntos de datos grandes, como aquellos con un gran número de campos de entrada.

El nodo KNN, análisis de vecino más próximo es un método de clasificación de casos basado en su similaridad con otros casos. En aprendizaje de máquinas, se ha desarrollado como una forma de reconocer patrones de datos sin requerir una coincidencia exacta con patrones o casos almacenados. Los casos parecidos están próximos y los que no lo son están alejados entre sí. Además, la distancia entre dos casos es una medida de sus diferencias.

El nodo de Análisis Discriminante realiza más supuestos rigurosos que las regresiones logísticas, pero puede ser una alternativa o un suplemento valioso al análisis de regresión logística si se cumplen dichos supuestos.

El nodo Red Bayesiana le permite crear un modelo de probabilidad combinando pruebas observadas y registradas con conocimiento del mundo real para establecer la probabilidad de las instancias. El nodo se centra en las redes Naïve Bayes aumentado a árbol (TAN) y de manto de Markov que se utilizan principalmente para la clasificación.

El nodo Lista de Decisiones identifica subgrupos, o segmentos, que muestran una mayor o menor posibilidad de proporcionar un resultado binario relacionado con la población global. Por ejemplo, puede buscar clientes que tengan menos posibilidades de abandonar o más posibilidades de responder favorablemente a una campaña. Puede incorporar su conocimiento empresarial al modelo añadiendo sus propios segmentos personalizados y previsualizando modelos alternativos, uno junto a otro para comparar los resultados. Los modelos de listas de decisiones constan de una lista de reglas en las que cada regla tiene una condición y un resultado. Las reglas se aplican en orden, y la primera regla que coincide determina el resultado.

La Regresión Logística es una técnica estadística para clasificar los registros en función los valores de los campos de entrada. Es análoga a la regresión lineal, pero toma un campo objetivo categórico en lugar de uno numérico.

El nodo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los nodos C&RT y QUEST, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar lo cálculos.

El nodo QUEST proporciona un método de clasificación binario para generar árboles de decisión; está diseñado para reducir el tiempo de procesamiento necesario para realizar los análisis de C&RT y reducir la tendencia de los métodos de clasificación de árboles para favorecer a las entradas que permitan realizar más divisiones. Los campos de entrada pueden ser continuos (rango numérico), sin embargo, el campo objetivo debe ser categórico. Todas las divisiones son binarias.

El nodo de Árbol de Clasificación y Regresión (C&R) genera un árbol de decisión que permite predecir o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera “puro” si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (sólo se crean dos subgrupos).

El nodo C5.0 genera un árbol de decisión o un conjunto de reglas. El modelo divide la muestra basándose en el campo que ofrece la máxima ganancia de información en cada nivel. El campo objetivo debe ser categórico. Se permiten varias divisiones en más de dos subgrupos.

El nodo Red Neuronal utiliza un modelo simplificado que emula el modo en que el cerebro humano procesa la información: Funciona simultaneando un número elevado de unidades simples de procesamiento interconectadas que parecen versiones abstractas de neuronas. Las redes neuronales son dispositivos eficaces de cálculo de funciones generales y requieren un conocimiento matemático o estadístico mínimo para entrenarlas o aplicarlas.

Nodos de modelado automático

Los nodos de modelado automático calculan y comparan un número de diferentes enfoques de modelado, facilitando la prueba de una variedad de métodos de una única pasada. Puede seleccionar los algoritmos de modelado que se utilizarán y las opciones específicas de cada uno de ellos, incluyendo combinaciones que de otro modo serían excluyentes entre sí. Por ejemplo, en lugar de elegir entre los métodos rápido, dinámico o de poda de una red neuronal, puede probarlos todos. El nodo explora cada combinación posible de opciones, evalúa cada modelo candidato en función de la medida especificada y guarda los mejores para su uso en la puntuación o en futuros análisis.

Puede seleccionar entre tres nodos de modelado automáticos, dependiendo de las necesidades de su análisis:

Nodos de modelado automático 1: Clasificador automático

El nodo Clasificador automático crea y compara varios modelos diferentes para obtener resultados binarios (sí o no, abandono o no de clientes, etc.), lo que le permite seleccionar el mejor enfoque para un análisis determinado. Son compatibles varios algoritmos de modelado, por lo que es posible seleccionar los métodos que desee utilizar, las opciones específicas para cada uno y los criterios para comparar los

resultados. El nodo genera un conjunto de modelos basado en las opciones especificadas y clasifica los mejores candidatos en función de los criterios que especifique.

Nodos de modelado automático 2: Autonumérico

El nodo Autonumérico calcula y compara modelos para resultados de rango numérico continúo utilizando cierto número de métodos diferentes. El nodo funciona de la misma manera que el nodo Clasificador automático, lo que le permite seleccionar los algoritmos que desee utilizar y experimentar con varias combinaciones de opciones en una única pasada de modelado. Los algoritmos admitidos incluyen redes neuronales, C&RT, CHAID, regresión lineal, regresión lineal generalizada y máquinas de vectores de soporte (SVM). Los modelos se pueden comparar basándose en la correlación, el error relativo o el número de variables utilizado.

Nodos de modelado automático 3: Agrupación en clústeres

El nodo Agrupación en clústeres automática calcula y compara los modelos de agrupación en clústeres que identifican grupos de registros con características similares. El nodo funciona de la misma manera que otros nodos de modelado automático, permitiéndole experimentar con múltiples combinaciones de opciones en una única pasada de modelado. Los modelos se pueden comparar utilizando medidas básicas con las que se intenta filtrar y definir la utilidad de los modelos de clúster y proporcionar una medida según la importancia de campos concretos.

Nugget de modelo

Un Nugget de modelo (Diamante) es el recipiente de un modelo, es decir, es el conjunto de reglas, fórmulas o ecuaciones que representan los resultados de las operaciones de generación de modelos en SPSS Modeler. La finalidad principal de un Nugget es puntuar datos para generar predicciones o permitir análisis adicionales de propiedades de modelos. Al abrir un Nugget de modelo en la pantalla, podrá ver diversos datos del modelo, como la importancia relativa de los campos de entrada en la creación del modelo.

Los mejores modelos se guardan en un único nugget de modelo compuesto, permitiendo explorarlos y compararlos y seleccionar los modelos que se utilizarán en la puntuación.

En objetivos numéricos, nominales y binarios únicamente, podrá seleccionar múltiples modelos de puntuación y combinar los resultados en un conjunto de modelos único. Al combinar predicciones de varios modelos, pueden evitarse las limitaciones en modelos individuales que suelen dar como resultado una precisión global superior que puede obtenerse de cualquiera de los modelos.

También puede decidir profundizar en los resultados y generar nodos de modelado o nugget de modelo para cualquier modelo individual que desee utilizar o explorar más a fondo.

IBM SPSS Modeler Server (Servidor)

Este componente es el encargado de leer los datos de diferentes fuentes, preparar los datos, modelar y comunicarse con el componente cliente. Por omisión el Componente Server de IBM SPSS MODELER convierte el Stream (Interface Visual del Usuario – Componente Cliente) en instrucciones SQL (Structured Query Languaje) para que la Base de Datos seleccione que operaciones pueden ser ejecutadas en SQL, las cree y las envíe al servidor de procesamiento. Las operaciones que la Base de Datos no pueda convertir en instrucciones SQL son ejecutadas en el Servidor de Procesamiento (IBM SPSS MODELER Server). El hecho que de varias operaciones puedan ser convertidas en sentencias SQL reduce el tráfico de datos por la red.

El componente Server lee los datos de diferentes fuentes y crea archivos temporales para aquellas operaciones que no son desarrolladas por la base de datos.

IBM SPSS Modeler Server ejecutara los procesos de extracción, transformación y preparación de datos aprovechando las capacidades de procesamiento en paralelo de la Base de Datos a través de la optimización automática de sentencias.

IBM SPSS Collaboration and Deployment (CAD)

IBM SPSS Collaboration and Deployment es una plataforma que permite manejar y automatizar el proceso analítico fácilmente y distribuir los resultados [54].

SPSS Enterprise Services permite:

- Centralizar y administrar sus activos analíticos para incrementar el conocimiento del negocio.
- Automatizar el proceso analítico para incrementar la productividad y asegurar resultados confiables, integrándolo con los demás procesos de negocio.
- Distribuir resultados con una interfaz personalizada a través de un browser.

Esta solución tiene 3 componentes: CAD Repository (Base), CAD Automation Services y CAD Real Time Scoring.

IBM SPSS CAD Repository:

Es la porción de solución que se encarga de trabajar en los inicios de sesión y mejorar la productividad en el análisis, puntos importantes los cuales garantizan la seguridad y el manejo correcto de la solución.

Puede almacenar rutas, nodos, modelos, paletas de modelos, proyectos y objetos de resultados en el repositorio, desde el que otros usuarios y aplicaciones pueden acceder a ellos.

También puede publicar un resultado de rutas en IBM® SPSS® Collaboration and Deployment Services Repository en un formato que permita a otros usuarios verlo en Internet mediante IBM® SPSS® Collaboration and Deployment Services Deployment Portal.

IBM SPSS CAD Automation:

Este módulo le permitirá hacer del análisis el componente principal del proceso de toma de decisiones de la organización. Pueden construirse procesos analíticos flexibles que pueden llevarse a la operación diaria con resultados consistentes. También le permite tener herramientas de gobernabilidad para administrar los diversos procesos de negocio

CAD Automation le permite automatizar tareas complejas de análisis, de tal manera que la organización pueda disponer permanentemente de los resultados, de acuerdo con sus necesidades.

Real Time Scoring Service

La arquitectura de SPSS Collaboration and Deployment Services se construye para apoyar un despliegue analítico de nivel empresarial. Al proporcionar un motor de datos común y la capa de abstracción de datos, la arquitectura se asegura de que los servicios de colaboración y distribución pueden ser integrados en entornos de aplicaciones múltiples, lo que permite a las organizaciones gestionar, de forma segura, el seguimiento de activos analíticos a través de su ciclo de vida

- Entregar los resultados analíticos que se producen con las interacciones con los clientes mediante la integración con los sistemas de los usuarios de negocios, y combinar la información recopilada durante la interacción con los datos históricos para determinar la puntuación.
- Incorporar características que aseguran la escalabilidad, fiabilidad y seguridad.
- Integración con aplicaciones existentes utilizando interfaces de programación estándar, tales como servicios Web.
- Soporte de aplicaciones de servidor de clusterización y virtualización para un uso más eficaz de los recursos.

Anexo 2: Procedimiento de calificación para la evaluación financiera

El procedimiento de calificación de los factores que se consideran para la evaluación financiera del potencial cliente, que en definitiva es la solvencia, comprende el determinar la condición que presenta o refleja el negocio en el momento de la solicitud de crédito, cada criterio dispone de diferentes niveles valorados por una puntuación que están parametrizados con anterioridad en la política de crédito del banco.

CARÁCTER: Este primer criterio se lo define como la integridad o el deseo de pagar de los clientes y es una variable cualitativa que se la mide en términos del grado de responsabilidad que el cliente evidencia. Por ser una variable difícil de cuantificar, se determinan las siguientes variables a ser consideradas y que serán puntuadas para obtener un valor cuantitativo en la calificación:

Comportamiento crediticio: Este aspecto tendrá el 50% de la ponderación general del carácter del cliente, dándole una calificación entre 1 y 10 que corresponderá al número de veces en que el cliente presentó retraso en sus pagos, según información revisada en el buró de crédito. Se asignará la siguiente puntuación:

Sector financiero

- Ningún retraso 7 puntos
- Menor o igual a 1 retraso 6 puntos
- Mayor o igual a 2 retrasos 3 puntos
- Mayor o igual a 3 retrasos 2 puntos
- Mayor o igual a 4 retrasos 0 puntos

Sector comercial

- Ningún retraso 3 puntos
- Menor o igual a 5 retraso 2 puntos
- Mayor o igual a 7 retrasos 1 puntos
- Mayor o igual a 8 retrasos 0 puntos

Calificación de Buró -Riesgos: La calificación asignada en el Buró de crédito tendrá una ponderación del 20% sobre la evaluación del carácter del cliente y se asignará una puntuación según los siguientes criterios:

Calificación 999 a 950 puntos 10 puntos

Calificación 949 a 900 puntos 8 puntos

Calificación 899 a 800 puntos 6 puntos

Calificación 799 a 750 puntos 5 puntos

Calificación 749 a 700 puntos 4 puntos

Calificación 699 a 600 puntos	2 puntos
Calificación 599 a 500 puntos	1 puntos
Calificación menor a 500	0 puntos

Entrevista: La entrevista in situ tendrá una ponderación del 10% y se asignará la puntuación respectiva tomando en consideración los siguientes criterios, donde prima la transparencia de la información que muestre el cliente y la congruencia entre lo que indica verbalmente y los soportes físicos que puedan visualizarse, desde el lugar de trabajo, las condiciones y entorno familiar.

Las respuestas del cliente fueron congruentes con su realidad y demostró transparencia en la documentación y/o información solicitada – 10 puntos

Las respuestas del cliente fueron congruentes con su realidad, pero no contaba con toda la documentación para demostrar cifras de años anteriores, pero se visualiza que lo informado está acorde a la realidad del negocio y forma de vida – 8 puntos

Las respuestas del cliente fueron convincentes con su realidad, pero no se aprecia orden en el manejo del negocio, pocos documentos de respaldo y se evidencia inexperiencia en el manejo del negocio – 5 puntos

Las respuestas del cliente fueron ambiguas, no se aprecia orden en el manejo del negocio, pocos documentos de respaldo y se evidencia inexperiencia en el manejo del negocio – 0 puntos

Los rangos de la puntuación van a ser medidos en base a la forma en cómo el Asesor de crédito /jefe de Agencia evidencia que el cliente maneja su negocio y economía familiar.

Zona geográfica: Este aspecto en el carácter del cliente será ponderado con el 20% y se considerará la siguiente evaluación, tomando en cuenta las zonas donde el Banco tiene presencia, infraestructura y demás condiciones de gestión para una adecuada atención de colocación y recuperación.

Clientes de la Zona Urbana	10 puntos
Clientes de la Zona Rural	5 puntos
Zonas vetadas o donde el Banco no tenga oficinas	0 puntos

Todas las ponderaciones serán relacionadas, es decir multiplicadas con el puntaje obtenido por el cliente para establecer una puntuación al Carácter del cliente.

CAPACIDAD: Este criterio considera la capacidad de pago como la primera fuente de repago y la experiencia en la administración del negocio.

El Asesor de crédito debe levantar la información necesaria para estructurar los estados financieros, los cuales permitirán determinar la capacidad de pago del cliente

considerando los ingresos y gastos de la microempresa, tomando en consideración los siguientes criterios que deberán ser ingresados en la FICHA DE EVALUACIÓN MICROEMPRESA la cual contendrá la metodología incorporada para una evaluación objetiva y que permita tener los criterios más adecuados para el análisis y calificación de crédito.

En este punto, se considerarán dos variables generales:

- **Situación financiera:** dentro de esta variable se considerarán los siguientes aspectos para la determinación de un puntaje, se obtendrá un indicador de relación entre ingresos versus los gastos del cliente (ingresos/gastos).
    - Dentro de los ingresos se considerarán tanto los ingresos generados por todos los conceptos por los que reciban un valor tanto por el solicitante y los de su cónyuge si aplica. (ingreso por ventas, comisiones adicionales, rentas)
    - En cuanto a los gastos, es importante determinar todos aquellos egresos que genere el cliente o su cónyuge si aplica (gastos de hogar, deuda financiera, donde incluya la cuota del microcrédito)
- **Destino del crédito:** En esta variable se asignará un puntaje tomando como consideración el uso que el cliente le dará al bien a adquirir, que en el caso de vehículos serán los siguientes:
    - Si el destino del crédito está enfocado específicamente en el giro del negocio - 10
    - Consumo para microempresarios – 8

La Situación financiera del cliente tendrá la ponderación del 80% y el destino del crédito corresponderá al 20% restante.

Todas las ponderaciones serán relacionadas, es decir multiplicadas, con el puntaje obtenido por el cliente tanto en la Situación financiera como en el Destino del crédito, para establecer una puntuación de la Capacidad de pago del cliente.

CAPITAL: Este criterio de capital tiene relación con la solvencia económica y financiera del solicitante. Se define como la diferencia entre el valor total de los activos menos el de las deudas y corresponde, por lo tanto, al nivel de inversión de fondos propios que el solicitante mantiene en su negocio, es decir, con los que podría destinar al pago de la deuda.

El Asesor de crédito/jefe de Agencia debe realizar el análisis del patrimonio por medio de indicadores, en especial el que mide la relación deuda patrimonio con el fin de evitar el sobreendeudamiento en el cliente.

El Capital se medirá bajo el concepto de Solvencia el cual tendrá una ponderación del 100% y que será puntuada de la siguiente manera:

- Patrimonio entre $0 y $1.000 4 puntos

- Patrimonio entre $1.001 y $2.500 5 puntos
- Patrimonio entre $2.501 y $5.000 6 puntos
- Patrimonio entre $5.001 y $10.000 6 puntos
- Patrimonio entre $10.001 y $20.000 8 puntos
- Patrimonio entre $20.001 y $30.000 9 puntos
- Patrimonio más de $30.000 10 puntos

La ponderación será relacionada, es decir multiplicada con el puntaje obtenido por el cliente para establecer una puntuación al Capital del cliente.

CONDICIONES: Se consideran Condiciones al entorno, es decir, la situación económica macro y micro, la situación del mercado, la situación política y las condiciones del préstamo, se tomarán en cuenta el análisis de las variables externas que pueden afectar el entorno, en el cual se desarrolla las actividades de la microempresa con la probabilidad de deteriorar en la capacidad de pago del cliente.

Para efectos del análisis del potencial cliente, el entorno será evaluado en función de los años que la empresa viene laborando, es decir, la situación laboral, a la cual se le asignará el 100% de esta variable y puntuada de la siguiente manera:

- Situación laboral menos de 1 año 2 puntos
- Situación laboral de 1 hasta 2 años 4 puntos
- Situación laboral de 2 a 4 años 7 puntos
- Situación laboral más de 4 años 10 puntos

La ponderación será relacionada, es decir multiplicada con el puntaje obtenido por el cliente para establecer una puntuación a la Condición del cliente.

COLATERAL: Esta variable se define como las garantías adecuadas y suficientes que respaldan el crédito, consideradas como la segunda fuente de repago de una obligación y en el caso de los microcréditos el colateral estará basado en una garantía real.

La ponderación para esta consideración será del 100% y equivaldrá a 10 puntos en la calificación total al tener como garantía real.

- Codeudor 10 puntos
- Garante 8 puntos
- Sin garante 6 punto

Anexo 3: Características de la Regresión Logística para créditos de consumo

Resumen de procesamiento de casos

		N	Porcentaje marginal
riesgoPago	NO	546	47,6%
	SI	601	52,4%
codigoEstadoCivilCliente	C	620	54,1%
	D	143	12,5%
	S	346	30,2%
	U	15	1,3%
	V	23	2,0%
tipoResidenciaCliente	A	431	37,6%
	F	315	27,5%
	N	372	32,4%
	P	26	2,3%
	S	3	0,3%
codigoProvinciaCliente	1	5	0,4%
	10	138	12,0%
	11	1	0,1%
	12	3	0,3%
	13	19	1,7%
	14	2	0,2%
	15	6	0,5%
	16	5	0,4%
	17	621	54,1%
	18	4	0,3%
	2	1	0,1%
	20	1	0,1%
	21	17	1,5%
	22	17	1,5%
	23	8	0,7%
	24	10	0,9%
	4	52	4,5%
	5	2	0,2%
	6	3	0,3%
	7	4	0,3%
	8	26	2,3%
	9	202	17,6%
codigoCiudadCliente		5	0,4%
	10.010000	85	7,4%
	10.020000	19	1,7%
	10.030000	6	0,5%
	10.040000	22	1,9%
	10.050000	4	0,3%
	10.060000	2	0,2%

Resumen de procesamiento de casos

	N	Porcentaje marginal
12.010000	1	0,1%
12.030000	1	0,1%
12.050000	1	0,1%
13.010000	7	0,6%
13.030000	2	0,2%
13.060000	1	0,1%
13.080000	7	0,6%
13.090000	1	0,1%
13.170000	1	0,1%
14.010000	1	0,1%
14.090000	1	0,1%
15.010000	2	0,2%
15.040000	3	0,3%
15.070000	1	0,1%
16.010000	5	0,4%
17.010000	537	46,8%
17.020000	59	5,1%
17.030000	3	0,3%
17.040000	14	1,2%
17.050000	3	0,3%
17.080000	3	0,3%
17.090000	2	0,2%
18.010000	3	0,3%
18.020000	1	0,1%
2.010000	1	0,1%
20.030000	1	0,1%
21.010000	13	1,1%
21.040000	3	0,3%
21.070000	1	0,1%
22.010000	16	1,4%
22.030000	1	0,1%
23.010000	8	0,7%
24.010000	3	0,3%
24.020000	5	0,4%
24.030000	2	0,2%
4.010000	38	3,3%
4.020000	5	0,4%
4.030000	4	0,3%
4.040000	1	0,1%
4.050000	4	0,3%
5.010000	2	0,2%
6.010000	3	0,3%

Resumen de procesamiento de casos

		N	Porcentaje marginal
7.010000		3	0,3%
7.070000		1	0,1%
8.010000		16	1,4%
8.040000		2	0,2%
8.050000		6	0,5%
8.060000		2	0,2%
9.010000		177	15,4%
9.040000		1	0,1%
9.060000		4	0,3%
9.070000		12	1,0%
9.090000		1	0,1%
9.100000		1	0,1%
9.110000		2	0,2%
9.160000		2	0,2%
9.210000		1	0,1%
9.250000		1	0,1%
claseVehiculo	Automóvil	563	49,1%
Bus		4	0,3%
Camión		73	6,4%
Camioneta		323	28,2%
Furgoneta		46	4,0%
Jeep		124	10,8%
Otros		14	1,2%
marcaVehiculo	BYD	2	0,2%
Changhe		1	0,1%
Chery		32	2,8%
Chevrolet		358	31,2%
Daihatsu		1	0,1%
DFSK		29	2,5%
Fiat		2	0,2%
Ford		17	1,5%
GreatWall		38	3,3%
Hino		1	0,1%
Hyundai		215	18,7%
Kia		140	12,2%
Mazda		107	9,3%
Mitsubishi		5	0,4%
Nissan		66	5,8%
Otros		5	0,4%
Peugeot		1	0,1%
Renault		40	3,5%
Suzuki		12	1,0%

Resumen de procesamiento de casos

		N	Porcentaje marginal
Toyota		58	5,1%
Wolkswagen		17	1,5%
tipoOperacion	COMERCIAL	6	0,5%
	CONSUMO	886	77,2%
	MICROCREDITO	255	22,2%
Válidos		1147	100,0%
Perdidos		0	
Total		1147	
Subpoblación		1147[a]	

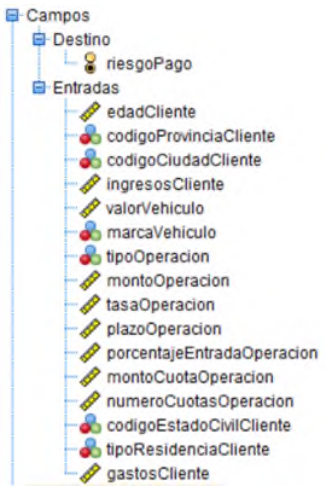
Información de ajuste de los modelos

Modelo	Criterios de ajuste de modelo	Pruebas de la razón de verosimilitud		
	Logaritmo de la verosimilitud -2	Chi-cuadrado	gl	Sig.
Sólo intersección Final	1587,441	856,362	108	,000
	731,079			

Pseudo R cuadrado

Cox y Snell	,526
Nagelkerke	,702
McFadden	,539

Anexo 4: Características del modelo con IA para créditos de consumo



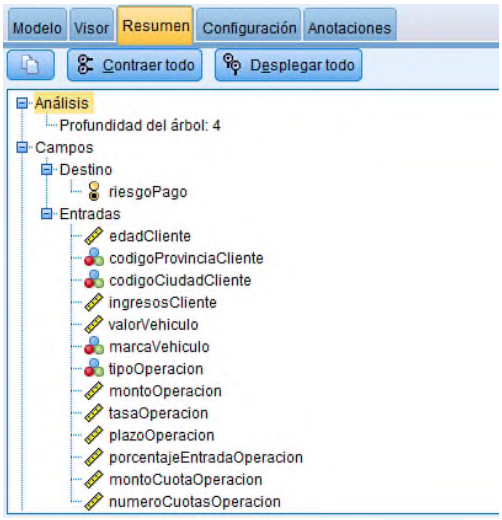
Configuración de creación

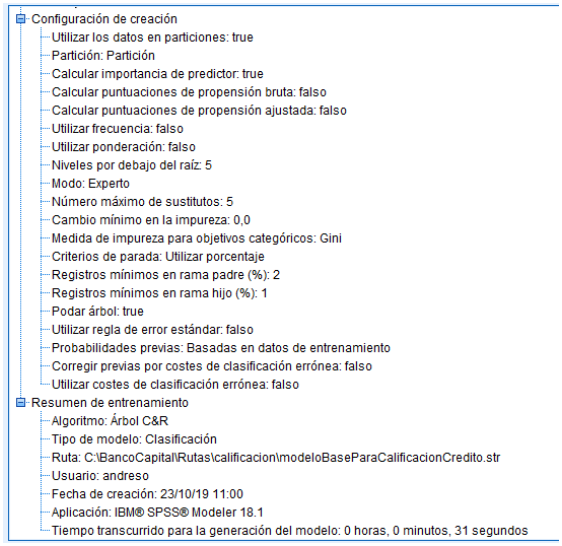
- Utilizar los datos en particiones: true
- Partición: Partición

Resumen de entrenamiento

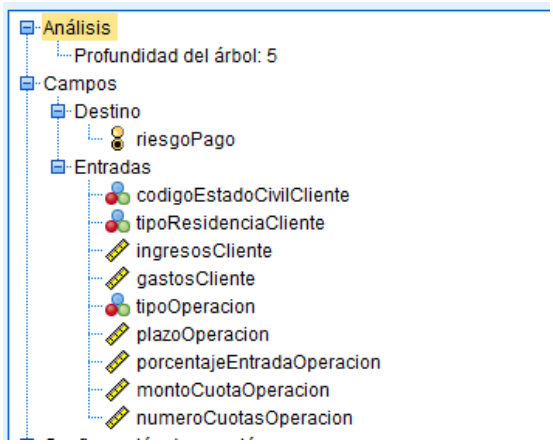
- Fecha de creación: 10/05/19 11:00
- Aplicación: IBM® SPSS® Modeler 18.1
- Modelos planificados: 13
- Modelos finalizados: 13
- Modelos descartados según los resultados finales: 10
- Modelos que no se han podido generar o puntuar: 0
- Modelos no finalizados debidos a una interrupción: 0
- Tiempo transcurrido para la generación del modelo: 0 horas, 0 minutos, 34 segundos

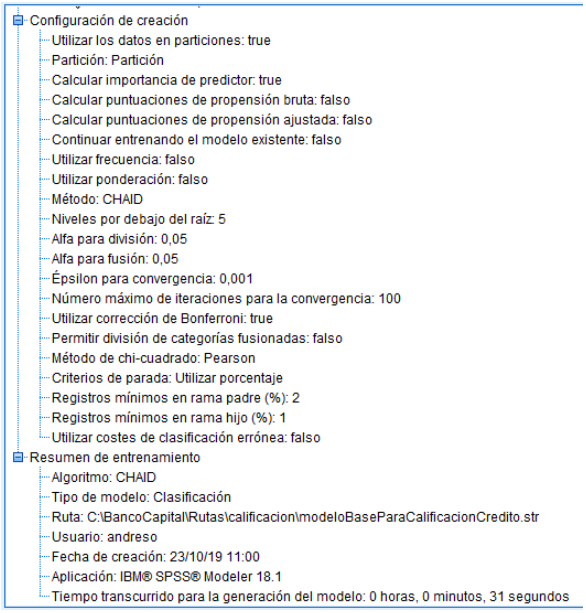
Árbol C&R 1



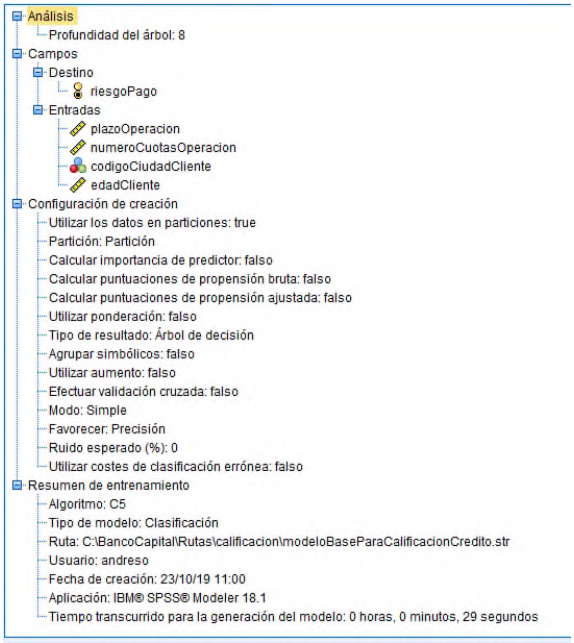


CHAID 1



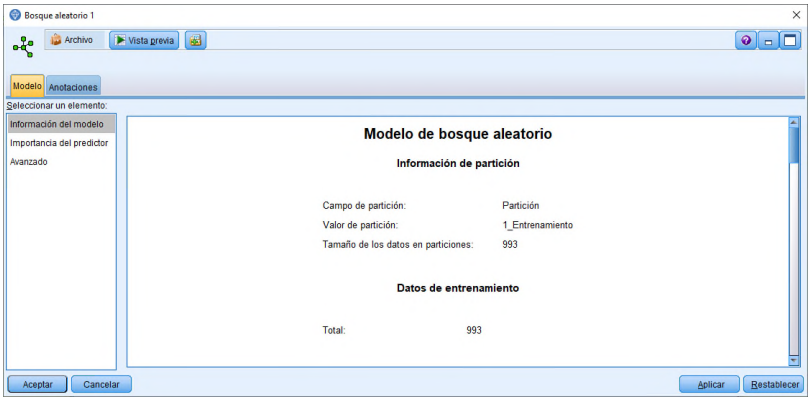

Configuración de creación
Utilizar los datos en particiones: true
Partición: Partición
Calcular importancia de predictor: true
Calcular puntuaciones de propensión bruta: falso
Calcular puntuaciones de propensión ajustada: falso
Continuar entrenando el modelo existente: falso
Utilizar frecuencia: falso
Utilizar ponderación: falso
Método: CHAID
Niveles por debajo del raíz: 5
Alfa para división: 0,05
Alfa para fusión: 0,05
Épsilon para convergencia: 0,001
Número máximo de iteraciones para la convergencia: 100
Utilizar corrección de Bonferroni: true
Permitir división de categorías fusionadas: falso
Método de chi-cuadrado: Pearson
Criterios de parada: Utilizar porcentaje
Registros mínimos en rama padre (%): 2
Registros mínimos en rama hijo (%): 1
Utilizar costes de clasificación errónea: falso
Resumen de entrenamiento
Algoritmo: CHAID
Tipo de modelo: Clasificación
Ruta: C:\BancoCapital\Rutas\calificacion\modeloBaseParaCalificacionCredito.str
Usuario: andreso
Fecha de creación: 23/10/19 11:00
Aplicación: IBM® SPSS® Modeler 18.1
Tiempo transcurrido para la generación del modelo: 0 horas, 0 minutos, 31 segundos

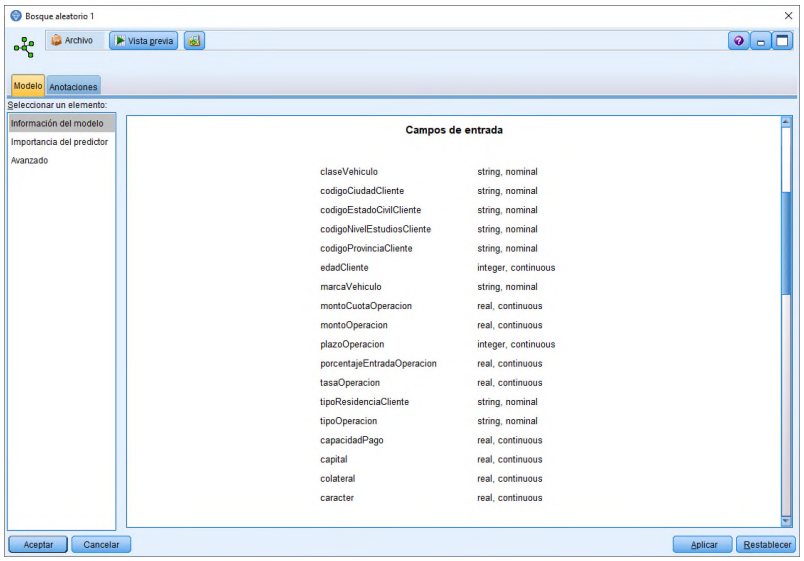
C5 1

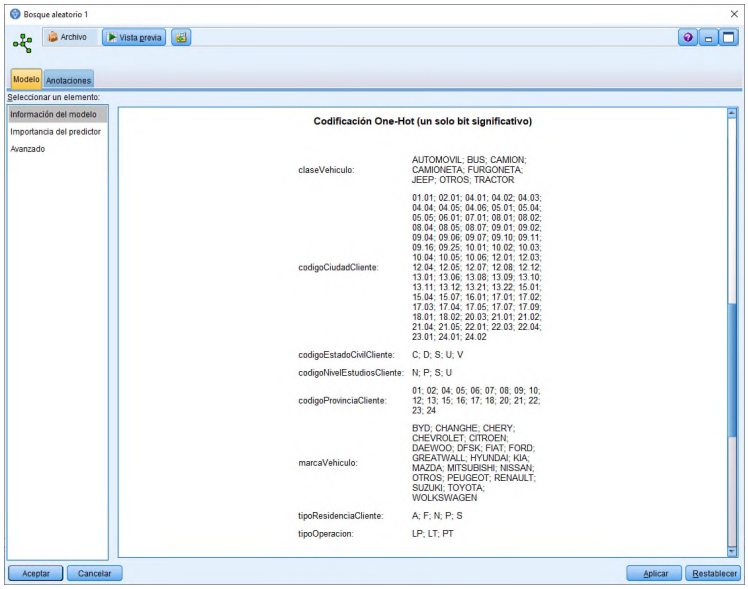

Análisis
Profundidad del árbol: 8
Campos
Destino
riesgoPago
Entradas
plazoOperacion
numeroCuotasOperacion
codigoCiudadCliente
edadCliente
Configuración de creación
Utilizar los datos en particiones: true
Partición: Partición
Calcular importancia de predictor: falso
Calcular puntuaciones de propensión bruta: falso
Calcular puntuaciones de propensión ajustada: falso
Utilizar ponderación: falso
Tipo de resultado: Árbol de decisión
Agrupar simbólicos: falso
Utilizar aumento: falso
Efectuar validación cruzada: falso
Modo: Simple
Favorecer: Precisión
Ruido esperado (%): 0
Utilizar costes de clasificación errónea: falso
Resumen de entrenamiento
Algoritmo: C5
Tipo de modelo: Clasificación
Ruta: C:\BancoCapital\Rutas\calificacion\modeloBaseParaCalificacionCredito.str
Usuario: andreso
Fecha de creación: 23/10/19 11:00
Aplicación: IBM® SPSS® Modeler 18.1
Tiempo transcurrido para la generación del modelo: 0 horas, 0 minutos, 29 segundos

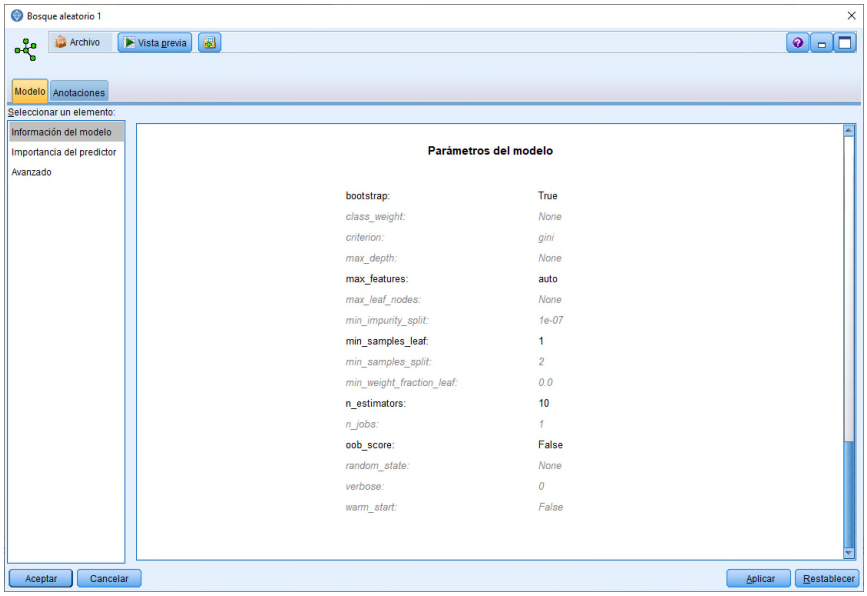
Anexo 5: Características del modelo diseñado con IA para microcréditos

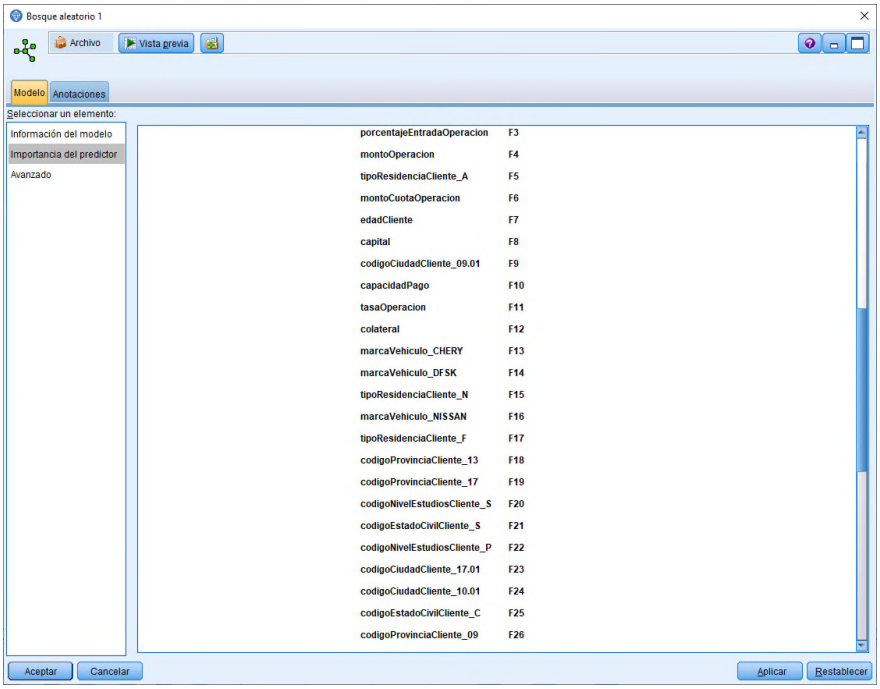
Modelo bosque aleatorio

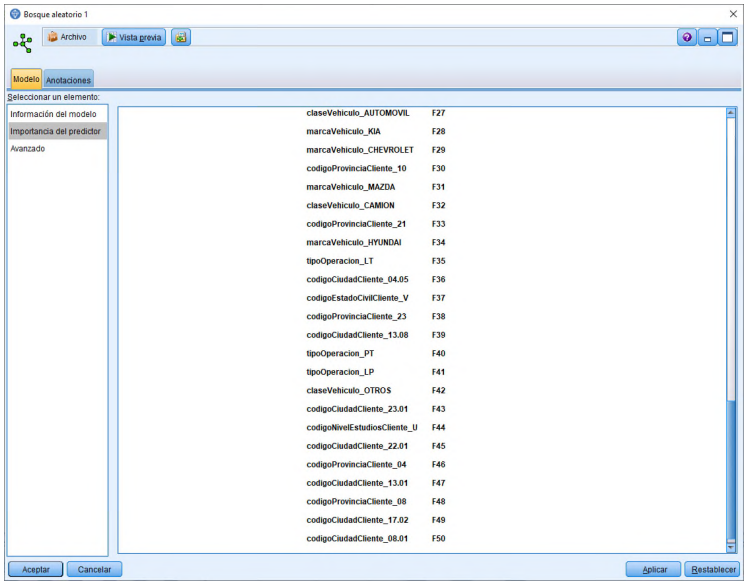




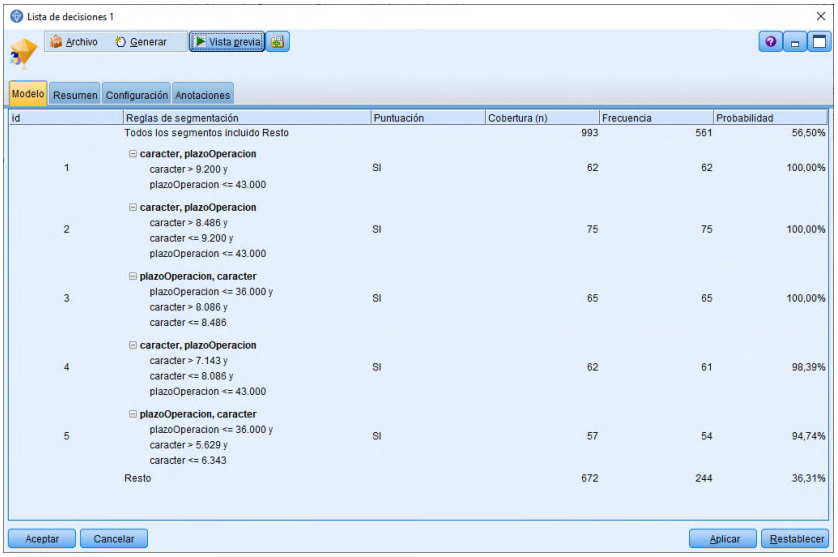


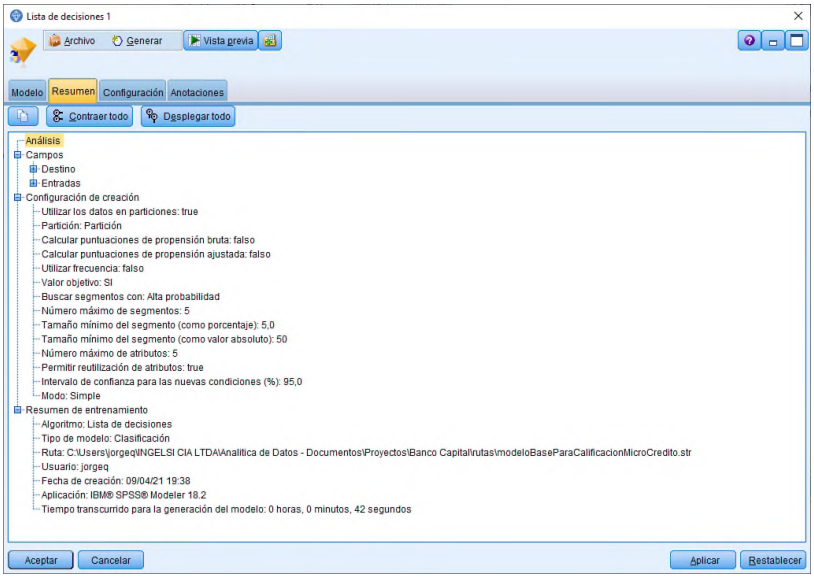

Bosque aleatorio 1
Archivo
Vista previa
Modelo
Anotaciones
Seleccionar un elemento:
Información del modelo
Importancia del predictor
Avanzado
Parámetros del modelo
bootstrap: True
class_weight: None
criterion: gini
max_depth: None
max_features: auto
max_leaf_nodes: None
min_impurity_split: 1e-07
min_samples_leaf: 1
min_samples_split: 2
min_weight_fraction_leaf: 0.0
n_estimators: 10
n_jobs: 1
oob_score: False
random_state: None
verbose: 0
warm_start: False
Aceptar
Cancelar
Aplicar
Restablecer


Bosque aleatorio 1
Archivo
Vista previa
Modelo
Anotaciones
Seleccionar un elemento:
Información del modelo
Importancia del predictor
Avanzado
porcentajeEntradaOperacion F3
montoOperacion F4
tipoResidenciaCliente_A F5
montoCuotaOperacion F6
edadCliente F7
capital F8
codigoCiudadCliente_09.01 F9
capacidadPago F10
tasaOperacion F11
colateral F12
marcaVehiculo_CHERY F13
marcaVehiculo_DFSK F14
tipoResidenciaCliente_N F15
marcaVehiculo_NISSAN F16
tipoResidenciaCliente_F F17
codigoProvinciaCliente_13 F18
codigoProvinciaCliente_17 F19
codigoNivelEstudiosCliente_S F20
codigoEstadoCivilCliente_S F21
codigoNivelEstudiosCliente_P F22
codigoCiudadCliente_17.01 F23
codigoCiudadCliente_10.01 F24
codigoEstadoCivilCliente_C F25
codigoProvinciaCliente_09 F26
Aceptar
Cancelar
Aplicar
Restablecer

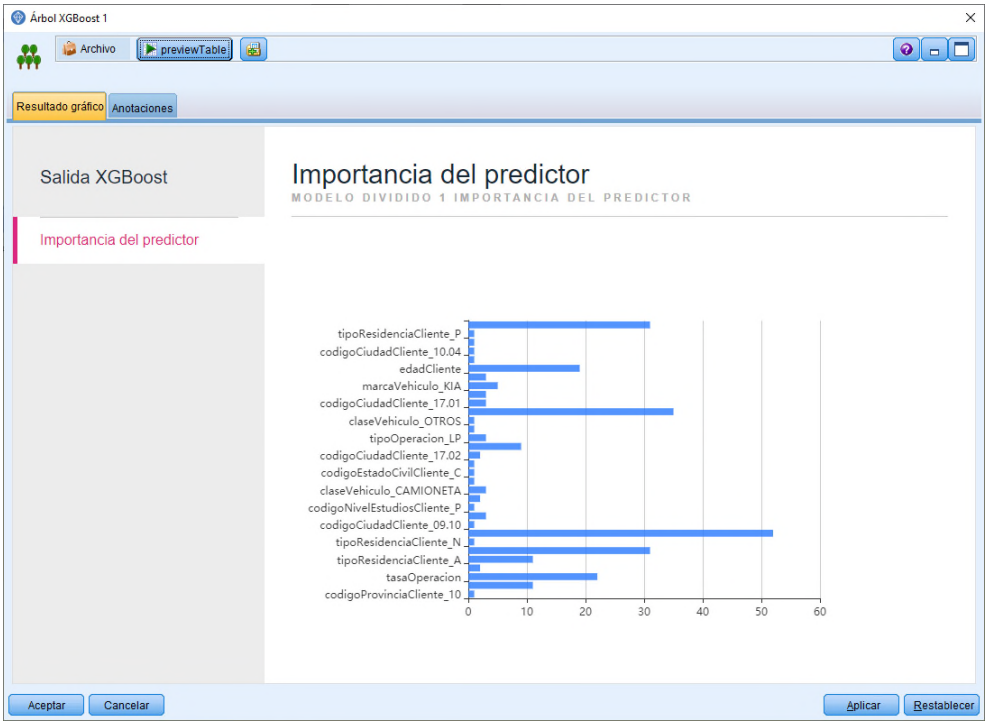


Modelo lista de decisiones

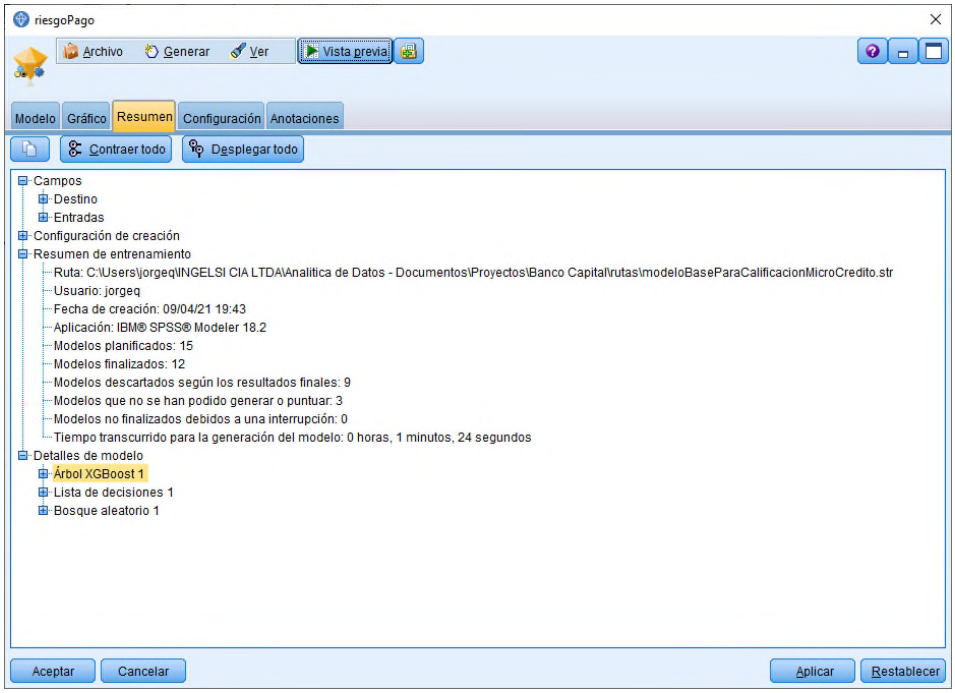




Modelo XGBoost



Resumen


riesgoPago
Archivo
Generar
Ver
Vista previa
Modelo
Gráfico
Resumen
Configuración
Anotaciones
Contraer todo
Desplegar todo
Campos
Destino
Entradas
Configuración de creación
Resumen de entrenamiento
Ruta: C:\Users\jorgeq\INGELSI CIA LTDA\Analitica de Datos - Documentos\Proyectos\Banco Capital\rutas\modeloBaseParaCalificacionMicroCredito.str
Usuario: jorgeq
Fecha de creación: 09/04/21 19:43
Aplicación: IBM® SPSS® Modeler 18.2
Modelos planificados: 15
Modelos finalizados: 12
Modelos descartados según los resultados finales: 9
Modelos que no se han podido generar o puntuar: 3
Modelos no finalizados debidos a una interrupción: 0
Tiempo transcurrido para la generación del modelo: 0 horas, 1 minutos, 24 segundos
Detalles de modelo
Árbol XGBoost 1
Lista de decisiones 1
Bosque aleatorio 1
Aceptar
Cancelar
Aplicar
Restablecer