UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA INDUSTRIAL Y DE SISTEMAS

SECCIÓN DE POSGRADO

TESIS:

“REDES NEURONALES ARTIFICIALES Y MÁQUINA CON SOPORTE VECTORIAL PARA CLASIFICAR A LOS SOLICITANTES DE MICROCRÉDITO”

PARA OBTENER EL GRADO ACADEMICO DE DOCTOR EN INGENIERÍA DE SISTEMAS

ELABORADA POR:

MAG. CÉSAR ALDO CANELO SOTELO

ASESOR:

DR. PEDRO CELINO ESPINOZA HARO

LIMA – PERÚ

2021

DEDICATORIA

A mi esposa Elba y a mis hijos, Isabel y Aldo, que son la motivación para seguir superándome.

AGRADECIMIENTO

A mi asesor, el Dr. Pedro Espinoza Haro por haber confiado en mí, por su paciencia y disposición a colaborar conmigo en la realización de esta investigación.

Al Dr. José Portillo Campbell y al Dr. Franco Krajnik Stulin por sus sugerencias valiosas para mejorar esta investigación.

RESUMEN

Las entidades crediticias constantemente se enfrentan al problema de controlar el riesgo de crédito al que se exponen al desarrollar sus operaciones crediticias, en tal sentido, siempre han requerido apoyarse en modelos predictivos que les ayuden a tomar decisiones acertadas para la aceptación o el rechazo de una solicitud de crédito. Los modelos predictivos que emplean las instituciones financieras para calificar a los solicitantes de crédito son los conocidos modelos clásicos basados en técnicas estadísticas y los modelos basados en técnicas de máquinas de aprendizaje.

En esta investigación, con el objetivo de clasificar a los solicitantes de microcrédito y contribuir a la mejora de la gestión del riesgo crediticio, se trabaja con la Base de datos de una Caja Municipal de Ahorro y Crédito (CMAC) que contiene 15,569 registros, cada uno con 27 variables, en donde las primeras 26 variables son los datos del cliente, y la última es la variable de aceptación o rechazo del crédito (V27). Al estudiar la Base de datos, se determinó que la variable Días de atraso de la última cuota pagada (V8) tiene la más alta correlación (0.78) con la Aceptación o el rechazo del crédito, y luego de un estudio más detenido de la Base de datos se descubrió que aquellos que tenían días de atraso de la última cuota superior a los 30 días se constituían en los clientes rechazados y los que no, en aceptados, por esta razón, las pruebas empleando las técnicas de máquinas de aprendizaje, se hacen con la Base de datos que, en unos casos incluyen esta variable y en otros casos la excluyen.

En primer lugar, se emplean las Redes Neuronales Artificiales (RNA) Backpropagation, para predecir el comportamiento crediticio de los prestatarios ante una entidad de microfinanzas. En segundo lugar, se emplean las Redes Self-Organizing-Maps (RNA-SOM) para agrupar los prestatarios en clústeres, y estudiar qué variables han influido en la conformación de los clústeres; y, en tercer lugar, se emplean las Máquinas con Soporte Vectorial (MSV) para separar los registros de la Base de datos.

Con la RNA Backpropagation se hicieron pruebas con diversas arquitecturas de la red, y se determinó que con una red de 4 capas con 14, 10, 8 y 1 neuronas respectivamente, se obtuvo una precisión 0.97682 que fue la mayor obtenida en todas las pruebas hechas con la Base de datos completa. Luego se han hecho pruebas con la Base de datos en la que la variable V8 sustituye a la variable V27, y la precisión obtenida es menor, pero sigue siendo una buena precisión de la red, y finalmente, se excluye de la Base de datos a la variable V8, obteniéndose una precisión menor, y de esta manera se comprueba que la variable V8 es la más realista.

Con Redes Self-Organizing-Maps (RNA-SOM) de dos neuronas se han hecho dos pruebas, una con topología Gridtop y métrica Dist, y otra prueba cambiando a la topología Hextop y a la métrica Linkdist. Los resultados obtenidos que están respaldados por las medidas estadísticas de las variables en cada clúster y los coeficientes de correlación en la formación de los clústeres, concluyen que, con el cambio de topología y métrica, no hay diferencias sustanciales en la composición de los clústeres, sólo ha habido el desplazamiento de un pequeño grupo de prestatarios entre los clústeres.

Con Redes Self-Organizing-Maps (RNA-SOM) de tres neuronas se han hecho dos pruebas, una con topología Gridtop y métrica Dist, y otra prueba cambiando a la topología Hextop y a la métrica Linkdist. Las medidas estadísticas de las variables en cada clúster y los coeficientes de correlación en la formación de los clústeres, permiten concluir que, con el cambio de topología y métrica, no hay diferencias sustanciales en la composición de los clústeres, sólo se ha producido el desplazamiento de un pequeño grupo de prestatarios entre los clústeres contiguos.

Finalmente, con la Máquina con Soporte Vectorial con núcleo lineal, se han separado a los prestatarios en dos grupos: aceptados y rechazados. Se han hecho dos pruebas, una con la Base de datos completa, y otra prueba eliminando la variable V8. En ambas pruebas se ha logrado la separación con un reducido número de vectores soporte en las fronteras, es decir se ha encontrado un hiper-plano de separación óptima que ha dado lugar a la separación de dos grupos de clientes bien definidos.

ABSTRACT

Credit institutions constantly face the problem of controlling the credit risk to which they are exposed when developing their credit operations, in this sense, they have always required to rely on predictive models that help them make the right decisions for the acceptance or rejection of a loan. credit request. The predictive models that financial institutions use to rate loan applicants are the well-known classical models based on statistical techniques and models based on machine learning techniques.

In this research, in order to classify microcredit applicants and contribute to improving credit risk management, we work with the Database of a Municipal Savings and Credit Fund (CMAC) that contains 15,569 records, each one with 27 variables, where the first 26 variables are the customer's data, and the last one is the credit acceptance or rejection variable (V27). When studying the Database, it was determined that the variable Days of arrears of the last installment paid (V8) has the highest correlation (0.78) with Acceptance or rejection of credit, and after a more detailed study of the Base of data, it was discovered that those who were days late in the last installment greater than 30 days became rejected customers and those who did not, accepted, for this reason, tests using machine learning techniques are made with the Database which, in some cases, includes this variable and in other cases they exclude it.

First, Backpropagation Artificial Neural Networks (ANNs) are used to predict the credit behavior of borrowers before a microfinance institution. Second, Self-Organizing-Maps Networks (RNA-SOM) are used to group borrowers into clusters, and study which variables have influenced the formation of the clusters; and, thirdly, the Vector Supported Machines (MSV) are used to separate the records from the Database.

With RNA Backpropagation, tests were carried out with various network architectures, and it was determined that with a 4-layer network with 14, 10, 8 and 1 neurons respectively, a precision of 0.97682 was obtained, which was the highest obtained in all the tests carried out. with the complete Database. Then tests have been done with the Database in which the variable V8 replaces the variable V27, and the precision obtained

is lower, but it is still a good precision from the network, and finally, the database is excluded from the variable V8, obtaining a lower precision, and in this way it is verified that the variable V8 is the most realistic.

With Self-Organizing-Maps Networks (RNA-SOM) of two neurons, two tests have been carried out, one with Gridtop topology and Dist metric, and another test changing to Hextop topology and Linkdist metric. The results obtained, which are supported by the statistical measures of the variables in each cluster and the correlation coefficients in the formation of the clusters, conclude that, with the change in topology and metric, there are no substantial differences in the composition of the clusters, there has only been the movement of a small group of borrowers between the clusters.

With Self-Organizing-Maps Networks (RNA-SOM) of three neurons, two tests have been done, one with Gridtop topology and Dist metric, and another test changing to Hextop topology and Linkdist metric. The statistical measures of the variables in each cluster and the correlation coefficients in the formation of the clusters, allow us to conclude that, with the change in topology and metric, there are no substantial differences in the composition of the clusters, only the displacement has occurred. of a small group of borrowers among the contiguous clusters.

Finally, with the Linear Core Vector Supported Machine, borrowers have been separated into two groups: accepted and rejected. Two tests have been done, one with the complete Database, and another test eliminating the variable V8. In both tests, separation has been achieved with a reduced number of support vectors at the borders, that is, an optimal separation hyperplane has been found that has resulted in the separation of two well-defined groups of clients.

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

La presente investigación está referida al problema de clasificar a los solicitantes de microcrédito en solicitudes aceptadas y rechazadas, con la finalidad de contribuir a la mejora de la gestión del riesgo crediticio, por tanto, el objeto de estudio es el solicitante de microcrédito. Este tipo de crédito se caracteriza por ser un préstamo destinado a fomentar la producción en microempresas o pequeños negocios que generan ingresos, con los cuales el prestatario mejora su nivel de vida y el de su familia. El acceso a este tipo de crédito es limitado, debido al riesgo de la no recuperación del crédito en el plazo pactado, las entidades crediticias restringen el otorgamiento de microcréditos, y los otorgan en condiciones más duras en cuanto a plazo, tasa de interés y garantías.

Para la investigación de este problema socioeconómico, se estudia la Base de datos de una microfinanciera compuesta por 15,569 registros de prestatarios. Con esta Base de datos se hacen diversas pruebas empleando técnicas de máquinas de aprendizaje para extraer conocimiento y hacer recomendaciones para contribuir a mejorar la gestión del riesgo crediticio de una microfinanciera.

En el capítulo I se presenta la realidad problemática con datos históricos y se define el problema de investigación, se justifica y se sustenta la importancia del problema de investigación, se plantean los objetivos y las hipótesis de la investigación. Así mismo se define el tipo y nivel de investigación.

En el capítulo II se revisa la literatura relacionada al tema de investigación, se desarrolla exhaustivamente el estado del arte referente a la temática tratada, y se presenta los trabajos previos de otros autores que abordan problemas afines, detallando las técnicas que utilizan y los resultados conseguidos. En este capítulo también se presenta el marco teórico del tema tratado.

En el capítulo III se estudia la Base de datos, se identifican las variables que tienen la mayor correlación con la variable de Aceptación o rechazo del crédito. En este capítulo también se hacen diversas pruebas con las Redes Neuronales Artificiales Backpropagation para predecir el comportamiento crediticio de los prestatarios, llegándose a determinar la arquitectura de la red con la cual se obtiene la mayor precisión de predicción. También se hacen pruebas con Redes SOM con dos y tres neuronas, variando la topología y la métrica, para determinar cómo la red agrupa a los prestatarios en clústeres. Por último, se hacen las pruebas con la Máquina con Soporte Vectorial con núcleo lineal para separar los prestatarios en aceptados y rechazados.

En el capítulo IV se presenta el análisis de las variables de la Base de datos, se la divide en dos sub-bases de datos, una de Aceptados y otra de Rechazados y se analizan las características de las variables que tienen la mayor correlación. Luego se analizan y discuten los resultados obtenidos de las pruebas hechas las Redes Neuronales Backpropagation, las Redes SOM, y la Máquina con Soporte Vectorial. Por último, se contrastan las hipótesis.

Finalmente se presentan las conclusiones y las recomendaciones.

CAPÍTULO I: METODOLOGÍA

1.1. DESCRIPCIÓN DE LA REALIDAD PROBLEMÁTICA

El riesgo crediticio es un problema antiguo que no ha cambiado en su esencia, pero se ha tornado cada vez más complejo por el crecimiento del volumen de las operaciones crediticias a través de los bancos. El riesgo crediticio es la posible pérdida ocasionada por el hecho de que un deudor incumpla con las obligaciones establecidas en el contrato de préstamo.

La gestión del riesgo crediticio es una tarea compleja para cualquier entidad financiera, y adquiere cada vez mayor importancia por las consecuencias sobre la liquidez y la estabilidad financiera que afectan a la entidad prestamista. Debido a que en todo crédito otorgado está inherente el riesgo de pérdida para el prestamista, las entidades crediticias en su interés de ser más competitivas, constantemente se enfrentan al problema de controlar el riesgo al que se exponen al desarrollar sus operaciones crediticias. Tanto las instituciones financieras y los organismos reguladores del sistema financiero coinciden en que la gestión del riesgo crediticio es un elemento esencial para el éxito a largo plazo.

Según Moradi y Mokhatab Rafiei (2019), la evaluación del riesgo de crédito es vital para los bancos; estos deben asegurarse de que los prestatarios puedan pagar sus cuotas antes de asignarles un préstamo. Según el Acuerdo de Basilea 2, cada banco necesita organizar y desarrollar su propio sistema interno de calificación crediticia con el que pueda analizar el riesgo de un prestatario.

Gulati, Goswami y Kumar (2018) se propusieron identificar las claves determinantes que impulsan el riesgo de crédito en la industria bancaria india durante el período 1998/1999 a 2013/2014. Destacan que, junto con las normativas y factores institucionales, hay una amplia gama de factores macroeconómicos y bancarios, específicos de la industria que influyen en la formación del riesgo de crédito bancario.

Afirman que la industria bancaria india ha experimentado una caída perceptible en la calidad de las carteras de los bancos desde 2010/2011. Las tasas brutas y netas de los préstamos en mora aumentaron a 4.9 y 2.7 por ciento del total de adelantos en 2014/2015, que es casi el doble de los niveles reportados en 2007/2008.

Ghosh (2018) que hace un estudio de sobre la morosidad en los bancos a través de sistemas de informes de crédito, afirma que los préstamos morosos constituyen un lastre para la actividad económica, especialmente para países donde los bancos son el pilar de la intermediación financiera. Señala tres consecuencias de la morosidad alta en los bancos:

- Disminución de la rentabilidad.
- Limitación del capital bancario.
- Aumento de los costos de financiación.

Dixon, Ritchie y Siwale (2007) afirman que en Zambia las instituciones de microfinanzas enfrentan niveles altos de morosidad e incumplimiento, altos costos operativos, todo lo cual restringen los esfuerzos para lograr la sostenibilidad financiera y organizacional que las ONG donantes requieren.

Según Barboza y Trejos (2009), para la viabilidad financiera de los programas de microcrédito, no solo es necesario mantener altas tasas de reembolso, sino también una baja morosidad, porque una alta tasa de morosidad crearía un efecto adverso en la disposición de los donantes para apoyar los programas de microcrédito. Los mercados financieros por su parte evitan los préstamos a pequeños prestatarios debido a que los préstamos son de monto reducido, tienen un alto riesgo de incumplimiento y un alto costo administrativo por dólar prestado.

Estas son algunas de las referencias a nivel global sobre el riesgo crediticio, todos estos autores coinciden en que las entidades crediticias deben mejorar el control

del riesgo de sus operaciones, y destacan la importancia que tiene el bajo índice de morosidad en la calidad de la cartera de créditos.

Con respecto a la morosidad en el Perú, las Estadísticas del Sistema Financiero que publica la Asociación de Bancos del Perú, Asbanc (2019), muestran que, a partir del 2012 hasta la actualidad, hay una tendencia creciente del índice de morosidad bancaria. En la Figura 1 se aprecia que en el 2012 el índice promedio de morosidad era de 1.70% y hasta agosto del 2019 la tasa promedio de morosidad es de 3.07%.

Figura 1. **Evolución del índice de morosidad en el Perú**

MOROSIDAD 1981 - 2019 (%)

	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Set	Oct	Nov	Dic
2000	9.36	9.95	9.81	10.21	10.28	10.05	10.49	10.47	10.34	10.34	10.21	9.98
2001	10.76	10.60	10.58	10.44	10.19	9.82	9.94	9.99	9.85	9.57	9.66	8.92
2002	9.38	9.05	9.00	8.82	8.60	8.02	8.35	8.31	8.09	8.15	8.35	7.58
2003	7.95	7.90	7.71	7.82	7.72	7.73	7.96	7.66	7.58	7.18	6.81	5.80
2004	5.98	5.79	5.75	5.69	5.50	5.10	5.06	4.93	4.59	4.47	4.14	3.71
2005	3.78	3.76	3.59	3.44	3.30	3.00	2.93	2.90	2.68	2.59	2.51	2.14
2006	2.26	2.29	2.10	2.10	2.08	2.00	2.03	1.93	1.86	1.87	1.81	1.63
2007	1.67	1.66	1.63	1.70	1.62	1.56	1.56	1.58	1.51	1.48	1.38	1.26
2008	1.40	1.38	1.36	1.31	1.31	1.21	1.22	1.21	1.19	1.19	1.26	1.27
2009	1.34	1.43	1.41	1.52	1.58	1.62	1.64	1.69	1.58	1.63	1.62	1.56
2010	1.66	1.67	1.73	1.72	1.76	1.66	1.81	1.75	1.64	1.63	1.59	1.49
2011	1.55	1.53	1.51	1.51	1.51	1.51	1.54	1.57	1.54	1.57	1.52	1.47
2012	1.54	1.60	1.62	1.71	1.72	1.73	1.72	1.75	1.72	1.79	1.79	1.75
2013	1.88	1.91	2.00	2.06	2.10	2.06	2.11	2.11	2.12	2.17	2.18	2.14
2014	2.28	2.30	2.34	2.37	2.45	2.36	2.44	2.46	2.41	2.47	2.46	2.47
2015	2.58	2.58	2.54	2.60	2.67	2.69	2.73	2.70	2.58	2.65	2.62	2.54
2016	2.64	2.71	2.70	2.77	2.86	2.87	2.85	2.91	2.86	2.95	2.96	2.80
2017	2.96	2.98	3.01	3.06	3.15	3.09	3.12	3.11	3.08	3.14	3.12	3.04
2018	3.12	3.24	3.07	3.11	3.14	3.10	3.18	3.23	3.07	3.10	3.07	2.95
2019	3.04	3.05	2.99	3.05	3.11	3.08	3.13	3.13				

Fuente: Asbanc 2018

Para América Latina, en la Figura 2, la misma fuente presenta las cifras bancarias que incluye la morosidad a junio de 2018 para 18 países, registrándose una tasa

promedio de morosidad de 2.5% para la región, siendo Colombia, con una tasa de 4.89%, el país que registra la mayor tasa morosidad. Perú registra una tasa de 3.10%, superior al promedio de los países de la región.

Figura 2. **Cifras bancarias en América Latina – junio 2018**

CIFRAS BANCARIAS AMÉRICA LATINA: Junio 2018

PAIS	ACTIVOS		COLOCACIONES		DEPOSITOS		PATRIMONIO		UTILIDAD		MOROSIDAD (%)	ROE (%)	INDICE DE SOLVENCIA (%)
	(MILLS. US$)	Moneda Local (Mills.)	(MILLS. US$)	Moneda Local (Mills.)	(MILLS. US$)	Moneda Local (Mills.)	(MILLS. US$)	Moneda Local (Mills.)	(MILLS. US$)	Moneda Local (Mills.)			
ARGENTINA	148,574	4,288,103	72,418	2,090,104	110,252	3,182,053	17,202	496,489	2,237	64,551	1.92	28.40	13.30
BOLIVIA	30,237	210,452	21,017	146,279	22,603	157,316	2,096	14,591	119	830	1.87	11.51	12.16
BRASIL	2,148,601	8,284,574	789,574	3,044,439	612,667	2,362,320	178,872	689,695	12,868	49,617	4.40	14.66	17.17
CHILE	361,318	234,116,063	257,066	166,565,792	205,197	132,957,332	29,258	18,957,471	2,003	1,298,000	1.93	13.69	13.04
COLOMBIA	200,798	588,499,999	146,379	429,008,789	123,200	361,073,286	26,015	76,245,480	1,456	4,268,553	4.89	11.51	15.76
COSTA RICA	39,676	22,606,829	24,793	14,126,684	27,238	15,520,101	3,947	2,249,071	122	69,516	2.76	6.18	14.01
ECUADOR	39,041	39,041	26,330	26,330	30,398	30,398	4,066	4,066	255	255	3.02	12.33	12.74
EL SALVADOR	17,457	17,457	12,149	12,149	12,184	12,184	2,227	2,227	81	81	1.97	7.61	16.16
GUATEMALA	41,596	311,686	23,362	175,056	30,530	228,770	3,905	29,259	287	2,150	2.36	17.46	14.56
HONDURAS	21,780	526,045	12,166	293,831	13,009	314,199	1,929	46,584	113	2,729	2.37	12.88	13.42
MEXICO	474,388	9,422,905	250,611	4,977,955	278,562	5,533,170	49,987	992,906	3,947	78,393	2.15	15.66	15.90
NICARAGUA	7,879	248,553	4,989	157,392	4,790	151,109	947	29,880	79	2,482	1.42	17.29	15.29
PANAMA	100,314	100,314	65,864	65,864	71,689	71,689	11,581	11,581	784	784	1.85	13.57	n.d.
PARAGUAY	20,931	119,362,897	13,344	76,098,179	15,353	87,554,476	2,411	13,746,637	232	1,322,092	3.03	23.32	10.75
PERU	112,219	367,181	78,344	256,340	71,015	232,362	13,307	43,540	1,235	4,041	3.10	18.70	15.07
REPUBLICA DOMINICANA	29,357	1,451,226	18,194	899,432	16,570	819,124	3,003	148,447	284	14,053	1.75	23.21	16.97
URUGUAY	35,577	1,119,470	16,237	510,911	28,420	894,278	3,250	102,275	408	12,836	4.01	25.10	n.d.
VENEZUELA	26,742	3,075,360,484	10,245	1,178,159,539	18,884	2,171,659,381	2,999	344,919,209	335	38,568,822	0.11	54.89	36.41

Fuente: Asbanc 2018

1.2. FORMULACIÓN DEL PROBLEMA

El problema es, dado una población de solicitantes de microcrédito, determinar quiénes serán potencialmente buenos clientes y quiénes serán potencialmente malos clientes, con la finalidad de contribuir a la mejora de la gestión del riesgo crediticio.

1.2.1. Problema general

¿De qué manera las técnicas de Máquinas de Aprendizaje contribuyen a predecir el comportamiento crediticio de los solicitantes de microcrédito, agruparlos y clasificarlos como aceptados o rechazados?

1.2.2. Problema específico

Problema Específico 1

¿Cuál es la base científica que soportan las Redes Neuronales Artificiales Backpropagation y cómo emplearlas para predecir el comportamiento crediticio del solicitante de microcrédito, y de esta manera, contribuir a la mejora de la gestión del riesgo crediticio?

Problema Específico 2

¿Cuál es la base científica que soportan las Redes Neuronales Artificiales *Self Organizing Maps* y cómo emplearlas para agrupar a los solicitantes de microcrédito, y de esta manera contribuir a la mejora de la gestión del riesgo crediticio?

Problema Específico 3

¿Cuál es la base científica que soportan las Máquinas con Soporte Vectorial y cómo emplearlas para clasificar a los solicitantes de microcrédito, como clientes aceptados o como clientes rechazados?

1.3. JUSTIFICACIÓN E IMPORTANCIA DE LA INVESTIGACIÓN

La justificación de la investigación se aborda destacando la importancia social, económica y financiera que tienen las microempresas, y en consecuencia los solicitantes de microcréditos en estos tres aspectos.

1.3.1. Importancia Social

Las microempresas juegan un papel clave en la cohesión social y son las principales generadoras de puestos de trabajo. Según los Estudios Económicos del Ministerio de la Producción, al año 2017, el 96.2% de las MIPYME eran microempresas,

y sus gestores son personas con iniciativa de negocio, que han demostrado tener un alto grado de correlación con el emprendimiento. En este contexto, se hace necesario el planteamiento y aplicación de políticas adecuadas que permitan y den soporte a su desarrollo.

Las microempresas son las principales usuarias de los microcréditos, y uno de los obstáculos que encuentran para su crecimiento y desarrollo, es la carencia de financiamiento para sus actividades. Por lo tanto, los microcréditos se constituyen como un instrumento adecuado para mejorar las condiciones de vida de las familias vulnerables, contribuir a su integración social y reducir la pobreza.

1.3.2. Importancia Económica

Las microempresas son importantes en la economía de cualquier país por su capacidad para generar empleo e ingresos, con lo cual fomentan el desarrollo y son un factor clave en la reducción de la pobreza.

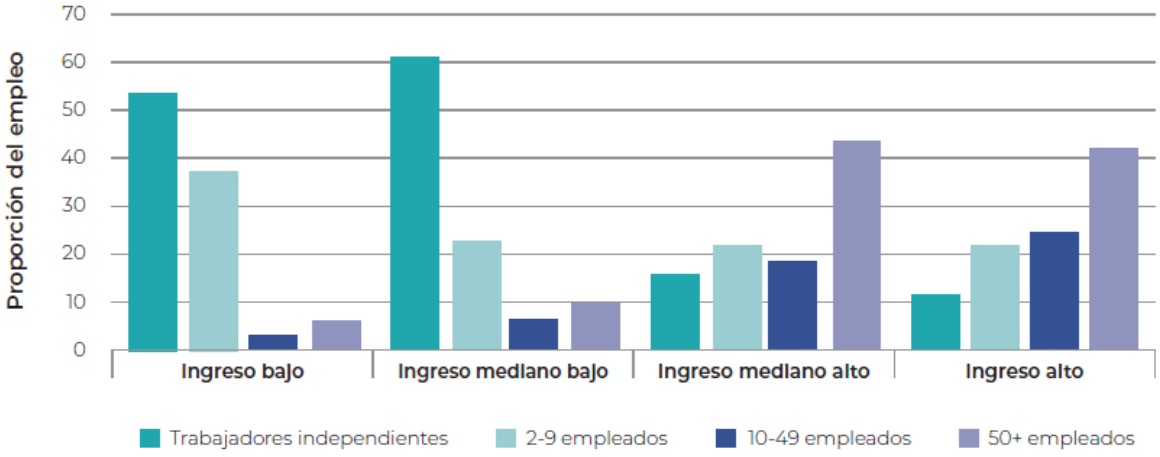
Un estudio realizado en el año 2019 por la Organización Internacional del Trabajo (OIT), hace estimaciones basadas en datos procedentes de encuestas nacionales de los hogares y de la población activa de 99 países de todas las regiones del mundo, excepto los países de América del Norte. Este informe proporciona datos empíricos a gran escala de gran utilidad sobre la contribución de los trabajadores independientes y de las empresas de distintos tamaños al total del empleo. De este estudio se obtienen las siguientes conclusiones:

a) A nivel global, las pequeñas unidades económicas conformadas por los trabajadores independientes, las microempresas y las pymes, representan la mayor parte del total del empleo.

b) Los emprendimientos, el empleo por cuenta propia y las microempresas representan el 70% del empleo total en 99 países.

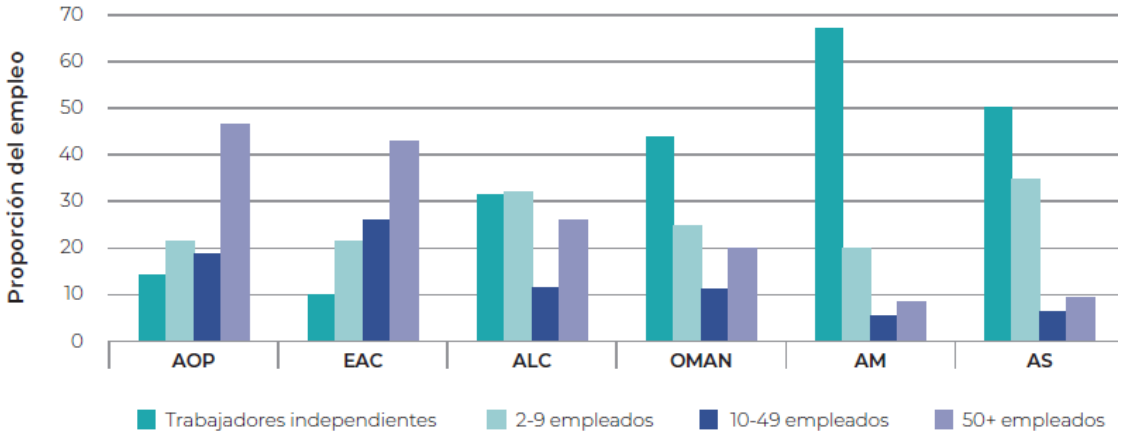
c) Hay una notable diferencia en el porcentaje de empleos generados por los trabajadores independientes y las microempresas en comparación con los empleos generados por la pequeña y mediana empresa, en los sectores de ingreso bajo e ingreso mediano bajo, tal como se aprecia en la Figura 3. Así mismo, hay regiones donde la proporción de empleo generado por los trabajadores independientes y las microempresas es notoriamente mayor que los empleos generados por las pymes, esto se muestra en la Figura 4.

Figura 3. **Proporción del empleo correspondiente a los trabajadores independientes y los diferentes tamaños de empresa, por nivel de ingreso.**



Fuente: Cálculos de la OIT, agosto de 2019.

Figura 4. **Proporción del empleo correspondiente a trabajadores independientes y a los distintos tamaños de empresa, por región.**



Nota: AOP = Asia Oriental y el Pacífico; EAC = Europa y Asia Central; ALC = América Latina y el Caribe; OMAN = Oriente Medio y África del Norte; AM = Asia Meridional; AS = África Subsahariana.

Fuente: Cálculos de la OIT, agosto de 2019.

d) La indiferencia en políticas públicas que ayuden a estos sectores a desarrollarse y salir adelante, el poco apoyo financiero, así como las condiciones de subsistencia de las micro y pequeñas empresas, dan paso a un mercado que emerge en la informalidad.

e) Cerca del 62% del empleo generado en estos 99 países estudiados por la OIT, corresponde al sector informal, donde las condiciones de trabajo en general tienden a ser inferiores, por la falta de seguridad social o salarios más bajos.

f) La forma de ayudar al sector no formal de la economía, a fin de incorporarlo plenamente al área productiva, es mediante servicios integrales y asesoría gerencial y técnica, crédito y asesoramiento en mercadeo. No se trata de ofrecer solidaridad al sector informal para que permanezca en el subdesarrollo; sino de impulsarlo hacia estadios superiores de la economía mediante instrumentos financieros, tecnológicos y administrativos adecuados.

g) El apoyo a las pequeñas unidades económicas debería ser una parte fundamental de las estrategias de desarrollo social y económico en todo el mundo, pero en particular en los países de ingreso bajo y mediano.

En lo que corresponde al Perú, los segmentos económicos que generan empleo y participan en la producción tienen una estructura similar a la que se tiene a nivel regional y global. Según los Estudios Económicos del Ministerio de la Producción, el segmento empresarial conformado por la micro, pequeña y mediana empresas (MIPYME) que operaban al 2017 sumaban más de 1.9 millones de empresas que representaban el 99.5% del total de empresas formales de la economía peruana. De ellas el 96.2% son microempresas, 3.2% pequeña y 0.1% mediana.

Parte significativa de la población y de la economía dependen de la actividad y el desempeño de este segmento debido a su reconocida capacidad para generar empleo y su participación en la producción. Las MIPYME generan alrededor del 60% de la PEA ocupada. Las microempresas dan empleo al 48% de la población ocupada.

En el período 2013-2017 el número de empresas formales de este segmento se ha incrementado a un ritmo promedio anual de 7.2%. Sin embargo, aún persiste un alto porcentaje de informalidad, ya que el 48.4% de las MYPE no están registradas en la SUNAT.

1.3.3. Importancia Financiera

Dada la importancia que tienen las microempresas en la generación de empleo y la promoción del crecimiento económico, es necesario apoyar su desarrollo, siendo uno de los aspectos claves el financiamiento de sus actividades. En ese sentido, los países en desarrollo deben promover la Inclusión Financiera, dado que uno de los principales obstáculos que afectan de manera desproporcionada a los pequeños negocios, es el acceso a la financiación.

La Inclusión Financiera se define como el acceso y uso de servicios financieros de calidad por parte de todos los segmentos de la población, que puede derivar en

importantes beneficios para el crecimiento económico y el bienestar general del país, además de contribuir a reducir la pobreza y a que todos tengan las oportunidades.

Un adecuado acceso a servicios financieros permite que la población prospere, amplíe sus oportunidades de consumo e inversión y mejore sus niveles de vida. Los individuos y empresas incluidos en el sistema financiero pueden administrar más fácilmente sus finanzas, su consumo y ahorro, iniciar o potenciar sus actividades productivas, gestionar sus riesgos y protegerse frente a eventos adversos. De este modo, la población incluida contribuye al crecimiento económico y a mejorar la competitividad y productividad del país; asimismo, aporta a la reducción de la pobreza y de las desigualdades y, trascendiendo lo económico, a su propio empoderamiento. Además, una mayor inclusión financiera contribuye a la estabilidad del sistema financiero, puesto que la mayor participación de la población ayuda a la diversificación de clientes, favoreciendo la estabilidad de las fuentes de recursos financieros y la calidad de la cartera de créditos.

En el Perú, según la Superintendencia de Banca, Seguros y AFP (SBS) sólo el 6% de las Mipyme acceden al sistema financiero regulado. La Tabla 1 contiene información de acceso al Sistema Financiero de las Mipyme, considerándose como acceso los créditos vigentes, vencidos, refinanciados y reestructurados utilizados en el año 2017. Se aprecia que las microempresas son el sector con más bajo acceso, pues solo el 4.6% accede al Sistema Financiero formal. Las pequeñas y medianas empresas muestran un mayor porcentaje de acceso.

Tabla 1. Acceso al Sistema Financiero año 2017

MIPYME	Número de empresas registrada en la Sunat	Número de empresas registradas en el Sistema Financiero diciembre 2017	Participación en el Sistema Financiero
Microempresa	1,836,848	83,839	4.6%
Pequeña	60,702	28,116	46.3%
Mediana	2,034	1,269	62.4%
Total	1,899,584	113,224	6.0%

Fuente: SBS 2017, Sunat 2017

Según el diario El Comercio (20 de noviembre de 2019), las investigaciones de la Unidad de Inteligencia de The Economist concluyen que el Perú es uno de los líderes en el entorno propicio para la inclusión financiera. Sin embargo, los retos para ayudar a reducir la desigualdad a través de la inclusión financiera son diversos. Según la Superintendencia de Banca, Seguros y Administradoras Privadas de Fondos de Pensiones (SBS), solo tres de cada diez adultos a nivel nacional tienen acceso al crédito. Por su parte el presidente de la Asociación de Instituciones de Microfinanzas (Asomif) comenta que se deberá enfatizar en la inclusión financiera productiva, para ayudar a las personas a salir de la pobreza y, así reducir las desigualdades.

En agosto de 2019, el Ejecutivo lanzó el Plan Nacional de Inclusión Financiera con una serie de metas al 2030 como: aumentar la participación de la población adulta con alguna cuenta en el sistema financiero a 75%; incrementar a 43% la participación de la población adulta con algún crédito con baja probabilidad de incumplimiento y elevar al 100% la cobertura del sistema financiero.

En el Decreto Supremo N° 255-2019-EF que aprueba la Política Nacional de Inclusión Financiera (profundización financiera y acceso al sistema financiero), se menciona que la inclusión financiera en el Perú es aún insuficiente y baja en relación con niveles internacionales, lo cual constituye un problema público, en tanto restringe la contribución del sistema financiero al crecimiento económico, a la productividad y competitividad y a la reducción de la pobreza y las desigualdades. Es innegable que se

han logrado avances importantes en los últimos años; por ejemplo, entre el 2015 y 2018, la proporción de adultos con al menos una cuenta en el sistema financiero se incrementó de 29% a 38% (ENAHO 2015 y 2018) y la correspondiente a créditos pasó de 30% a 33% entre el 2013 y 2018 (SBS, 2018). No obstante, estas cifras revelan que un gran porcentaje de la población, concentrada principalmente en los segmentos más vulnerables, no tiene participación en el sistema financiero.

Entre los factores que limitan la inclusión financiera están la pobreza y la informalidad, que son dos grandes problemas estructurales del país. Si bien estos factores limitan la inclusión financiera, la mejora de esta última contribuye a la reducción de dichos problemas. Respecto a la informalidad, la tasa de informalidad laboral estimada al cierre del 2017 fue de 72.5% (INEI, 2018), lo que constituye una cifra elevada en términos absolutos y en comparación con niveles internacionales. Por otro lado, si bien la pobreza en el Perú ha venido disminuyendo de manera importante en los últimos años, al cierre del 2018, el 20.5% de la población se encontraba en situación de pobreza y el 2.8% en situación de pobreza extrema (INEI, 2019). La informalidad y la pobreza están asociadas, entre otros aspectos, a reducidos niveles de ingreso y productividad, a una preferencia por el uso de dinero en efectivo y a escasa información, todo lo cual limita el acceso a los servicios financieros.

A pesar de los múltiples beneficios que conlleva la inclusión financiera para las personas, las empresas, el propio sistema financiero y la economía en su conjunto; las cifras ponen en evidencia la existencia de un problema público. Si bien se han logrado avances, en el Perú todavía existen bajos niveles de acceso y uso de servicios financieros de calidad por parte de la población, en relación a otros países de la región. Así, la proporción de personas mayores de quince años con al menos una cuenta en el sistema financiero se incrementó de 20% en el 2011 a 43% en el 2017 (Base de Datos de Global Findex de 2011 y 2017)[1], todavía por debajo del promedio registrado en

[1] En el 2011, el Banco Mundial puso en marcha Global Findex, la base de datos más completa del mundo sobre las modalidades que usa la gente para ahorrar, pedir préstamos, realizar pagos y gestionar riesgos. Global Findex abarca más de 140 economías del mundo. A la encuesta inicial en 2011, le siguieron una segunda en el 2014 y una tercera en el 2017. La Base de datos de 2017 se hizo en base a encuestas a más 150,000 adultos de más de 140 economías de todo el mundo.

América Latina (55%) en el 2017. Asimismo, el porcentaje de adultos con un crédito del sistema financiero pasó de 26% al 33% entre el 2011 y 2018.

Por otra parte, el uso de los instrumentos de pagos digitales para la realización de transacciones es reducido. En el 2017, el 34% de la población hizo o recibió algún pago a través de medios digitales, mientras que en América Latina dicho porcentaje alcanzaba el 46% (Base de Datos de Global Findex, 2017).

A nivel mundial, el Informe de Global Findex 2017, muestra que la inclusión financiera está aumentando, y que 1,200 millones de adultos han abierto sus cuentas desde el 2011, esto significa que el 69% de los adultos de todo el mundo, posee una cuenta, en comparación con el 62% en 2014 y el 51% en el 2011.

Concluimos que un sector financiero desarrollado contribuye a movilizar y redistribuir los recursos, así como a gestionar el riesgo crediticio, lo que favorece el crecimiento del sector privado. La financiación fomenta el crecimiento económico, que a su vez genera empleo.

1.4. OBJETIVOS

1.4.1. Objetivo General

Determinar modelos basados en técnicas de Máquinas de Aprendizaje para predecir el comportamiento crediticio de los solicitantes de microcrédito, agruparlos y clasificarlos como aceptados o rechazados.

1.4.2. Objetivos Específicos

Objetivo Específico 1

Conocer la base científica de los modelos de Red Neuronal Artificial Backpropagation y seleccionar la arquitectura más adecuada, para predecir el comportamiento crediticio de los solicitantes de microcrédito con la más alta precisión.

Objetivo Específico 2

Conocer la base científica de las Redes Self Organizing Maps para determinar un modelo que agrupe convenientemente a los solicitantes de microcrédito.

Objetivo Específico 3

Conocer la base científica de las Máquinas con Soporte Vectorial para determinar un modelo que clasifique a los solicitantes de microcrédito como aceptados o rechazados.

1.5. HIPÓTESIS

1.5.1. Hipótesis General

Modelos basados en técnicas de Máquinas de Aprendizaje predecirán el comportamiento crediticio de los solicitantes de microcrédito, los agruparán y clasificarán como aceptados o rechazados en función de todas las variables involucradas en la Base de Datos.

1.5.2. Hipótesis Específica

Hipótesis Específica 1

Un modelo de Red Neuronal Artificial Backpropagation con la arquitectura adecuada, predecirá el comportamiento crediticio de los solicitantes de microcrédito con la más alta precisión.

Hipótesis Específica 2

Un modelo basado en Red Self Organizing Maps agrupará convenientemente a los solicitantes de microcrédito.

Hipótesis Específica 3

Un modelo basado en Máquina con Soporte Vectorial clasificará a los solicitantes de microcrédito como aceptados o rechazados.

1.6. VARIABLES E INDICADORES

En cuanto al número de variables a considerar en el modelo, Rayo, Lara y Camino (2010) en su investigación sobre un modelo de Credit scoring para una institución de microfinanzas en el Perú, consideran 40 variables que son agrupadas en variables del cliente, variables de la operación de crédito y variables del ciclo económico. Por su parte, Baklouti (2013) en su investigación agrupa las variables en tres categorías: variables sociodemográficas, variables características de préstamos y variables conductuales. A su vez, Bekhet y Eletter (2014) en su investigación trabajan con trece variables, siete de ellas son variables de escala y seis son categóricas. Blanco et al. (2013) agrupan las variables en cinco categorías: características personales, ratios financieros y económicos de la microempresa, características de la operación financiera actual y retraso en el pago del microcrédito. Por su parte Nanayakkara y Stewart (2015) en su estudio sobre los determinantes para el reembolso en las microfinanzas en Indonesia y Sri Lanka consideran 16 variables que los agrupan en características del prestatario, características del préstamo y características de la entidad crediticia.

Teniendo en cuenta estos antecedentes, en el presente estudio se trabajará con una Base de datos con registros que contienen variables características del préstamo y variables conductuales del prestatario.

1.7. MATRIZ DE CONSISTENCIA

Los problemas, objetivos, hipótesis y variables se definen en la matriz de consistencia que se muestra en la Tabla 2. Esta matriz permite evaluar el grado de coherencia y conexión lógica entre el problema, objetivos, las hipótesis y las variables.

Tabla 2. Matriz de Consistencia

PROBLEMA GENERAL	**OBJETIVO GENERAL**	**HIPÓTESIS GENERAL**	**VARIABLES**
¿De qué manera las técnicas de Máquinas de Aprendizaje de contribuyen a predecir el comportamiento crediticio de los solicitantes de microcrédito, agruparlos y clasificarlos como aceptados o rechazados?	Determinar modelos basados en técnicas de Máquinas de Aprendizaje para predecir el comportamiento crediticio de los solicitantes de microcrédito, agruparlos y clasificarlos como aceptados o rechazados.	Modelos basados en técnicas de Máquinas de Aprendizaje predecirán el comportamiento crediticio de los solicitantes de microcrédito, los agruparán y clasificarán como aceptados o rechazados en función a todas las variables involucradas en la Base de datos.	X = Variables características del préstamo, variables conductuales del prestatario. Y = Resultado del modelo basado en Máquinas de Aprendizaje.
PROBLEMAS ESPECÍFICOS	**OBJETIVOS ESPECÍFICOS**	**HIPÓTESIS ESPECÍFICAS**	**VARIABLES**
1) ¿Cuál es la base científica que soportan las Redes Neuronales Artificiales Backpropagation y cómo emplearlas para predecir el comportamiento crediticio del solicitante de microcrédito, y de esta manera, contribuir a la mejora de la gestión del riesgo crediticio?	1) Conocer la base científica de un modelo de Red Neuronal Artificial Backpropagation y seleccionar la arquitectura más adecuada, para predecir el comportamiento crediticio de los solicitantes de microcrédito con la más alta precisión.	1) Un modelo de Red Neuronal Artificial Backpropagation con la arquitectura adecuada, predecirá el comportamiento crediticio de los solicitantes de microcrédito con la más alta precisión.	X = Variables características del préstamo, variables conductuales del prestatario. Y = Resultado del modelo predictivo.

2) ¿Cuál es la base científica que soportan las Redes Neuronales Artificiales Self Organizing Maps y cómo emplearlas para agrupar a los solicitantes de microcrédito, y de esta manera contribuir a la mejora de la gestión del riesgo crediticio?	2) Conocer la base científica de las Redes Self Organizing Maps para determinar un modelo que agrupe convenientemente a los solicitantes de microcrédito.	2) Un modelo basado en Red Self Organizing Maps agrupará convenientemente a los solicitantes de microcrédito.	X = Variables características del préstamo, variables conductuales del prestatario. Y = Clústeres que agrupan a los prestatarios.
3) ¿Cuál es la base científica que soportan las Máquinas con Soporte Vectorial y cómo emplearlas para clasificar a los solicitantes de microcrédito, como clientes aceptados o como clientes rechazados, y de esta manera contribuir a la mejora de la gestión del riesgo crediticio?	3) Conocer la base científica de las Máquinas con Soporte Vectorial para determinar un modelo que clasifique a los solicitantes de microcrédito como aceptados o rechazados.	3) Un modelo basado en Máquina con Soporte Vectorial clasificará a los solicitantes de microcrédito como aceptados o rechazados.	X = Variables características del préstamo, variables conductuales del prestatario. Y= Clasificación de los solicitantes de créditos.

Fuente: Elaboración propia

1.8. UNIDAD DE ANÁLISIS

La unidad de análisis es el prestatario de microcrédito. Es una persona natural o jurídica que generalmente solicita un microcrédito para desarrollar una actividad productiva.

1.9. TIPO Y NIVEL DE INVESTIGACIÓN

El tipo de la presente investigación se define a partir de los siguientes criterios:

- Por el objetivo de estudio es una investigación aplicada, porque se busca encontrar conocimientos que se puedan aplicar para resolver problemas prácticos.

- Por el método de estudio de las variables es una investigación cuantitativa, pues se han obtenido datos numéricos para las variables estudiadas.

- Por el número de variables es una investigación multivariada, debido a que se estudian un conjunto de variables independientes y una variable dependiente.

- Por la naturaleza del problema la investigación es no experimental, porque las variables independientes no pueden ni deben ser manipuladas.

- Por el período de tiempo en que se realiza el estudio, es una investigación transeccional o transversal, dado que se recolectan los datos en un solo momento en el tiempo.

La investigación es de nivel correlacional-explicativo-predictivo, pues trata de medir la relación existente entre dos o más variables, explicar las características de las variables y pronostica el comportamiento crediticio del solicitante de microcrédito.

1.10. DELIMITACIÓN DE LA INVESTIGACIÓN

La investigación está limitada al estudio del riesgo crediticio de los solicitantes de microcrédito. Este tipo de crédito es otorgado a las personas naturales y a las microempresas para actividades productivas. Las entidades crediticias otorgan estos microcréditos en condiciones desventajosas debido al mayor riesgo inherente a este tipo de crédito.

1.11. FUENTES DE INFORMACIÓN E INSTRUMENTOS UTILIZADOS

Dado que la investigación estará limitada al estudio del riesgo de los solicitantes de microcrédito, la población estará constituida por las carteras de crédito de entidades que otorgan microcrédito. Estas entidades están constituidas por:

- Entidades bancarias privadas
- Entidades financieras privadas
- Cajas Rurales de Ahorro y Crédito (CRAC)
- Cajas Municipales de Ahorro y Crédito (CMAC)
- Entidades de Desarrollo de la Pequeña y Microempresa (EDPYMES)
- Cooperativas de Ahorro y Crédito
- La muestra será la cartera de microcréditos de una Caja Municipal de Ahorro y Crédito.

1.12. TÉCNICAS O PROCEDIMIENTOS DE RECOLECCIÓN Y PROCESAMIENTO DE DATOS

La recolección de datos consistirá en la obtención de una Base de datos de la cartera de microcréditos de una entidad microfinanciera. Los datos serán procesados empleando técnicas de Máquinas de aprendizaje.

CAPÍTULO II: MARCO TEÓRICO

El tema de estudio es el empleo de modelos basados en técnicas de máquinas de aprendizaje para la predicción del comportamiento crediticio de los solicitantes de microcrédito, agruparlos convenientemente y clasificarlos como buen riesgo o mal riesgo.

Para la presente investigación se ha tratado de identificar las investigaciones relacionadas con el tema de estudio, que de acuerdo a la taxonomía de la IEEE corresponde a Computational and Artificial Intelligence\Prediction Methods\Predictive models.

2.1. REVISIÓN DE LA LITERATURA

Se hace una selección de los artículos publicados en revistas de los bancos: Science Direct, Scopus, Springer, Taylor and Francis y Ebsco, en el período comprendido entre los años 2005 y 2020. Se fija como inicio del período el año 2005 porque a partir de ese año, los modelos y métodos propuestos se centraron más en el uso de la inteligencia artificial.

Para la búsqueda de la información en los artículos se han empleado las palabras claves: Microcredit, Credit Scoring, Credit Risk y Data Mining. Estas palabras deben mencionarse en al menos alguna de las partes del artículo científico: el título, el resumen y las palabras claves. La selección se ha ido ajustando, como criterio de inclusión se seleccionaron los documentos relacionados a la predictibilidad del solicitante de microcrédito, excluyéndose las actas de congresos.

De la revisión de la literatura, se ha identificado que los modelos predictivos emplean diversos métodos de predicción. Nurlybayeva y Balakayeva (2013) hacen un análisis de diferentes técnicas de modelado de calificación crediticia que pueden utilizarse para el procesamiento de grandes conjuntos de datos, describen los métodos y tecnologías básicas del desarrollo de modelos de puntuación para la gestión de riesgos

del sistema bancario, y agrupan los métodos de predicción en estadísticos y no estadísticos. Por su parte, Dželihodžić, Đonko y Kevrić (2018) hacen una categorización más amplia de estas técnicas, y los clasifican en los siguientes grupos:

a) *Clasificadores individuales:* este grupo representa modelos de puntuación de crédito que solo utilizan una sola técnica de clasificación en el modelo, estos a su vez se subdividen en dos subgrupos:

   - Técnicas estadísticas
   - Técnicas de máquinas de aprendizaje (machine learning)

b) *Clasificadores de conjunto:* múltiples clasificadores como una combinación de clasificadores individuales para mejorar el rendimiento de la clasificación. Un clasificador de conjunto consiste en un conjunto de clasificadores entrenados individualmente, llamados clasificadores base, cuyas decisiones se combinan de alguna manera, típicamente mediante votación ponderada o no ponderada al clasificar nuevos ejemplos.

c) *Clasificadores híbridos:* son una combinación de dos o más técnicas heterogéneas de máquina de aprendizaje.

Con ese criterio se revisa la literatura y clasificamos los modelos de puntuación de crédito en tres grupos.

2.1.1. Clasificadores individuales

a) Técnicas Estadísticas

Rayo et al. (2010) desarrollan un modelo de credit scoring basado en la Regresión Logística Binaria para una entidad especializada en micro créditos en el Perú. Este modelo está basado en el conocimiento de las características del préstamo en el momento de su desembolso y su comportamiento de pago después del desembolso. Se emplean la Regresión Logística por su flexibilidad en el tratamiento de las variables

categóricas, y porque esta técnica permite determinar la influencia de cada variable independiente sobre la variable dependiente. El modelo obtenido es capaz de predecir correctamente el 78.3% de los créditos de la cartera de la microfinanciera.

Baklouti (2013) aborda objetivamente los factores determinantes relacionados con las tasas de reembolso de las micro-finanzas, para contribuir a la mejora del rendimiento de las amortizaciones a través de una mejor exploración de sus determinantes. Trabaja con una muestra de préstamos concedidos por el Microfinance Bank de Túnez en el período 2001-2009, y emplea un modelo de Regresión Logística Binaria para determinar la probabilidad de reembolso del prestatario. Los resultados obtenidos le permiten identificar un grupo de variables que han tenido un efecto notable en la estrategia de reembolso del préstamo.

Nanayakkara et al. (2015) desarrollaron un modelo basado en la Regresión Logística para identificar los factores determinantes para predecir el éxito de las amortizaciones de los microcréditos en los países de Sri Lanka e Indonesia. El modelo encuentra diferencias significativas entre los dos países. Se identifica 7 variables significativa en Sri Lanka con lo cual el modelo alcanza 76% de precisión, y en Indonesia identifica 3 variables significativas con lo que el modelo alcanza un 70% de precisión en la predicción.

Abid, Masmoudi y Zouari-Ghorbel (2016) identifican el modelo de calificación crediticia con mejor desempeño para los bancos tunecinos. Para este fin, se utiliza la Regresión Logística, así como el análisis Discriminante para desarrollar modelos predictivos que distinguen entre “buenos” y “malos” prestatarios. Trabaja con una base de datos de un banco tunecino que incluía clientes nuevos y existentes durante el período 2010-2012. Empleó un modelo de regresión logística para identificar la probabilidad de que cada solicitud pertenezca a una clase específica. Usó Análisis Discriminante con el propósito de maximizar la diferencia entre dos grupos, mientras que las diferencias entre los miembros particulares del mismo grupo son minimizadas. Los resultados muestran que el modelo de Regresión Logística clasifica correctamente el 99% de las observaciones de la muestra, y el modelo de Análisis Discriminante clasificó

correctamente a sólo el 68.49% de los solicitantes. Por tanto, hay una superioridad del modelo de Regresión Logística en comparación del análisis discriminante en términos de predicción de pagos los prestatarios incumplidos.

b) Técnicas de Máquinas de Aprendizaje

Zhou, Lai y Yu (2008) desarrollaron un modelo de credit scoring basado en la Máquina con Soporte Vectorial con mínimos cuadrados (LSSVM) y con la finalidad de optimizar el modelo, los parámetros fueron ajustados empleando el método de búsqueda directa (DS). Para los experimentos y posterior comparación, se emplearon otros métodos de selección de parámetros: Grid search (GD), Diseño de experimentos (DOE) y Algoritmos genéticos (GA). Se usaron dos conjuntos de datos del mundo real y para medir la eficiencia de los modelos se usaron la sensibilidad, la especificidad y la precisión general. De los resultados obtenidos, el modelo propuesto de búsqueda directa DS-LSSVM tuvo la mejor precisión general entre los cuatro modelos, en los dos conjuntos de datos. Posteriormente, el modelo DS-LSSVM se comparó con otros cinco clasificadores: Análisis Discriminante Lineal (LDA), Análisis Discriminante Cuadrática, Regresión Logística (LogR), Árboles de decisión (DT) y k-vecino más cercano (k-NN); y el modelo propuesto muestra también la mejor precisión general en ambos conjuntos de datos.

Dereliog y Gürgen (2011) proponen un método para el análisis del riesgo crediticio basado en una Red Neuronal Perceptron Multicapa para las pequeñas y medianas empresas de Turquía. El método propuesto incluye la selección de características usando árboles de decisión (DT) y la extracción recursiva de características con máquina de soporte vectorial (RFE-SVM), luego la extracción de características mediante el análisis de factores (FA) y análisis de componentes principales (PCA), y finalmente se usan los clasificadores k-vecinos más cercanos (k-NN), redes neuronales perceptron multicapa (MLP) y máquina con soporte vectorial (SVM). Los resultados obtenidos muestran que MLP es el clasificador de mejor rendimiento en términos de precisión de la predicción.

Khashman (2010) utiliza modelos de Redes Neuronales supervisada Backpropagation para la evaluación del riesgo crediticio bajo diferentes sistemas de aprendizaje. Las redes neuronales son entrenadas utilizando datos de solicitudes de crédito alemán. Usa tres modelos de redes neuronales que difieren en su topología, luego se investigan nueve esquemas de aprendizaje con diferentes porcentajes de datos para el entrenamiento y la validación, y se comparan los resultados de la implementación. Para comparar los tres modelos de redes neuronales dentro de los nueve esquemas de aprendizaje se emplean los criterios tasa de precisión y costos computacionales. Solo tres implementaciones satisfacen estrechamente los criterios de evaluación.

Zhou, Jiang, Shi y Tian (2011) propusieron un nuevo método de aprendizaje basado en kernel denominado Kernel afín al subespacio de los puntos más cercanos (KASNP) para la evaluación del crédito. En comparación con la Máquina con Soporte Vectorial (SVM), el método KASNP evita la solución del problema de programación cuadrática convexa, porque es un problema sin restricciones y calcula directamente la solución óptima para el conjunto de entrenamiento. Este método se prueba con tres conjuntos de datos de crédito del mundo real, para la evaluación y comparación con SVM se usa la precisión de la predicción, y los resultados experimentales indican que es más efectivo y competitivo.

Khashman (2011) utiliza el nuevo enfoque de Redes Neuronales Emocionales (EmNNs) y compara su rendimiento con las Redes Neuronales convencionales (NNs) en la evaluación del riesgo crediticio. Afirma que, a pesar de la aplicación exitosa de las NNs en la evaluación de una solicitud de crédito, esta red no puede entregar un resultado determinante para la toma de decisiones, porque la red neuronal carece de los factores emocionales que son la ansiedad y la confianza. Emplea un conjunto de datos de crédito de Australia y se entrenan 12 redes neuronales (seis modelos convencionales y seis emocional) utilizando tres diferentes sistemas de aprendizaje. Para evaluar el rendimiento se emplean los criterios número de iteraciones para la formación de la red neuronal, valor de error mínimo requeridos, tiempo de funcionamiento de la red

entrenada y tasa de precisión. El modelo de red neuronal EmNN-1 supera a las 11 redes neuronales restantes en los cuatro criterios de evaluación.

Bekhet y Eletter (2012) desarrollan un modelo predictivo de alto rendimiento usando Redes Neuronales Artificiales (ANN) para los bancos comerciales jordanos. Trabajan con un conjunto de datos de préstamos de bancos comerciales jordanos. Construyen una Red Neuronal feed-forward de tres capas con la arquitectura 12-9-1 y se utiliza el método de entrenamiento por lotes para reducir el error más rápidamente. Obtiene como resultado que el modelo neuronal podría detectar el 95% de las solicitudes aceptadas correctamente, y el 89.9% de las solicitudes rechazadas correctamente.

Shi, Zhang y Qiu (2013) proponen un modelo para la calificación crediticia empleando Máquina con Soporte Vectorial y Bosques aleatorios para la ponderación de las características (RF-FWSVM). Este modelo se compara con los modelos F-score con SVM con características ponderadas (FS-FWSVM) y SVM clásico, para lo cual se usan dos conjuntos de datos del mundo real. La evaluación del rendimiento del modelo se hace mediante el porcentaje correctamente clasificado y la desviación estándar de la precisión de la clasificación. Los resultados obtenidos indican que el modelo propuesto tiene el mejor rendimiento en las medidas de evaluación.

Blanco, Pino-Mejías, Lara y Rayo (2013) desarrollan un modelo de credit scoring para la industria de microfinanzas utilizando Redes Neuronales Perceptron Multicapa (MLP) y comparan su rendimiento con tres técnicas paramétricas: Análisis Discriminante Lineal (LDA), Análisis Discriminante Cuadrática (QDA) y Regresión Logística (RL). Trabajan con una base de datos de una microfinanciera peruana que contiene información de clientes de microcrédito del período 2003-2008. Se crean 17 modelos de credit scoring, de los cuales 14 modelos se basan en MLP. Para evaluar los modelos se emplea el área bajo la curva ROC (AUC) y los costos de clasificación errónea. De los resultados obtenidos se concluye que, los modelos MLP no solo tienen un AUC mayor, sino también tienen un costo de clasificación errónea menor.

Han, Han y Zhao (2013) se proponen mejorar la calificación de créditos a través de la reducción de dimensiones, mediante la Regresión Logística y la Máquina con Soporte Vectorial (SVM), definiendo la “reducción de la dimensión ortogonal” (ODR). Para enfrentar el problema de la alta dimensión debido a la multicolinealidad entre las variables, utilizan la hibridización con la Regresión Logística (HLR) para la selección de características y el Análisis de Componentes Principales (PCA) para la extracción de características. Experimentan con diversos modelos, y concluyen que ningún modelo puede ser completamente superior a otros modelos, pero el modelo HLG-ODR-SVM tiene la ventaja de la reducción ortogonal de dimensiones resultando ser el más efectivo para el problema de credit scoring, debido a que tiene la mayor precisión en la predicción de los malos créditos.

Beltran, Muñoz y Muñoz (2014) construyen un clasificador eficiente usando Redes Bayesianas, de mayor precisión que otros modelos para el problema de credit scoring. Se trabaja con un conjunto de datos de clientes que solicitaron un crédito a una caja de ahorros. El enfoque bayesiano basado en modelos de probabilidad muestra una capacidad predictiva superior respecto a los resultados obtenidos por los métodos Regresión Logística, Máquinas con Soporte Vectorial, Redes Neuronales, Árbol de clasificación y métodos multiclasificadores.

Bekhet y Eletter (2014) exploran la eficacia de dos modelos de calificación de crédito en los bancos comerciales jordanos: Red Neuronal con función de base radial (RBF) y el modelo de Regresión Logística. También se investiga la superioridad del modelo RBF sobre la Regresión Logística en la identificación de los potenciales morosos. Trabajan con un conjunto de solicitudes aceptadas y rechazadas de diferentes bancos comerciales jordanos del período 2006-2011. El modelo de regresión logística tiene como finalidad determinar la probabilidad de que un solicitante pertenece a una clase, bueno o malo. Esta investigación también mide la importancia que tienen las variables para el modelo. Ambos modelos muestran resultados prometedores y se concluye que no hay un mejor modelo general de evaluación de una solicitud de crédito. El modelo de regresión logística obtiene mejor resultado en la tasa de clasificación general, pero el

modelo RBF lo supera en la selección de solicitudes rechazadas, la identificación de potenciales morosos y por tanto la minimización del error tipo II.

Malhotra y Malhotra (2014) primero evalúan la efectividad de las Redes Neuronales y las técnicas estadísticas para detectar potenciales morosos de préstamos en el entorno de una cooperativa de crédito. Segundo, investigan la superioridad de varios modelos de Redes Neuronales existentes sobre técnicas estadísticas. Tercero, comparan la precisión de la clasificación de tres modelos de Redes Neuronales para evaluar las solicitudes de préstamos al consumidor: El modelo Backpropagation con aprendizaje adaptativo, el modelo Backpropagation con la aproximación de Levenberg-Marquardt y el modelo de Cuantificación de vectores de aprendizaje (aprendizaje competitivo) y seleccionar el mejor modelo que identifica el máximo de préstamos "malos". Trabajan con las solicitudes de préstamos de cooperativas. El modelo Backpropagation con aproximación Levenberg-Marquardt con dos capas ocultas tiene mejor rendimiento que los demás modelos, pues tiene la mayor precisión de los errores tipo II que es el más costoso para las entidades crediticias. El modelo de cuantificación del vector de aprendizaje (aprendizaje competitivo) proporcionó el grado más alto de exactitud global de la predicción en la identificación correcta de "buenos" préstamos y "malos" préstamos de crédito.

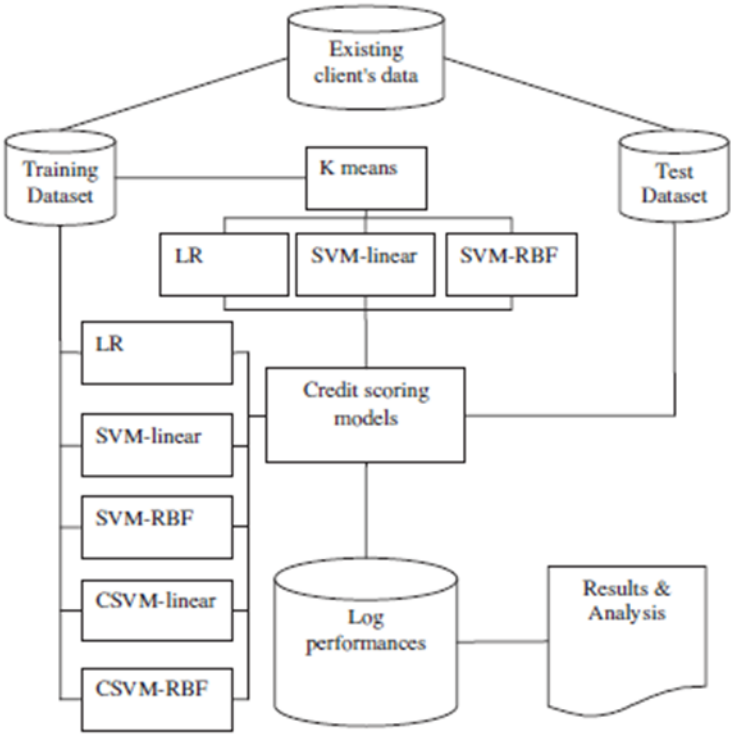
Zhao, Xu, Kang, Kabir, Liu y Wasinger (2015) presentan un modelo de credit scoring de mejor precisión basado en Redes Neuronales Feed forward perceptron multicapa (MLP). La mejora del modelo se basa en tres aspectos: optimizar la distribución de los datos en el conjunto de datos utilizando el nuevo método denominado Elección aleatoria promedio (para garantizar la equidad y el equilibrio de los datos), comparar los efectos del número de instancias en las fases de entrenamiento, validación y prueba, y encontrar el número adecuado de neuronas en la capa oculta. Se utiliza un conjunto de datos del mundo real y los resultados muestran que este modelo logra una precisión de clasificación de 87% que es superior a la de los modelos reportados en la literatura relevante de los últimos años.

Kiruthika y Dilsha (2015) comparan un modelo de Regresión Logística y un modelo de Red Neuronal para la calificación crediticia. Trabajan con un conjunto de datos de la India consistente en 520 registros de microcréditos. Para la evaluación de los modelos emplearon la tasa de clasificación errónea y el área bajo la curva ROC (AUC), ambos modelos se probaron con y sin selección de variables. Los resultados indicaron que la selección de variables influye en el rendimiento del modelo, por tal razón el modelo de Red Neuronal con selección de variables resultó ser el mejor. Otra conclusión del estudio indica que no hay un parámetro específico y una regla para construir un buen modelo ya sea para Red Neuronal o Regresión Logística, cada modelo tiene sus ventajas y desventajas y ambos modelos requieren decisiones cuidadosas para especificar su arquitectura.

Harris (2015) diseña un algoritmo usando Máquina con Soporte Vectorial agrupado (CSVM) para la calificación crediticia. Dado que los conjuntos de datos históricos de puntuación de crédito son grandes, los enfoques no lineales son altamente precisos y computacionalmente costosos, este estudio compara la CSVM con otras técnicas no lineales basadas en SVM con núcleo y muestra que el CSVM puede alcanzar niveles comparables de rendimiento de clasificación mientras permanece relativamente barato computacionalmente. Los datos fueron pre-procesados para transformar todas las categorías en datos numéricos para su análisis. Además, los datos fueron normalizados para mejorar el rendimiento del CSVM y los otros siete clasificadores desarrollados como comparadores. Se desarrollaron los siguientes clasificadores: Regresión Logística, K means más Regresión Logística, Máquina con Soporte Vectorial agrupado con un kernel RBF, K means más la Máquina con Soporte Vectorial con un kernel RBF, Máquina con Soporte Vectorial con un kernel RBF, Máquinas con Soporte Vectorial lineales agrupados, K means más una Máquina con Soporte Vectorial con un núcleo lineal, y una Máquina con Soporte Vectorial lineal. La Figura 5 presenta una vista de alto nivel de los algoritmos implementados. Para la construcción del modelo, cada muestra del conjunto de datos se dividió aleatoriamente en dos grupos de datos: prueba (20%) y entrenamiento y validación cruzada (80%). El conjunto de datos de prueba se utilizó exclusivamente para probar el desempeño de los modelos de clasificación

desarrollados. Este enfoque da cierta intuición en cuanto al rendimiento de los modelos en la configuración del mundo real. El conjunto de datos de entrenamiento y validación cruzada se utilizó para desarrollar los modelos para cada tipo de clasificador. En términos de rendimiento los modelos CSVM (tanto lineales como RBF) mostraron un área bajo la curva ROC (AUC) comparables a los otros clasificadores, ya que no hubo diferencias significativas entre ellos y los otros clasificadores en términos de AUC.

Figura 5. **Vista de alto nivel del sistema**



Fuente: Tomado de Terry (2014)

Los resultados indican que los modelos lineales CSVM superan constantemente a sus comparadores directos.

Leong (2015) propone un modelo de Red Bayesiana para la calificación del riesgo crediticio. Para los experimentos usa un conjunto de datos del mundo real y las medidas de evaluación que emplea son exactitud, la sensibilidad, la precisión y el área bajo la curva ROC. El rendimiento del modelo es comparado con el de los modelos de Regresión Logística y Redes Neuronales en cuatro muestras de datos que se diferencian por su tamaño y por el equilibrio o no de los datos que contienen. Los resultados obtenidos indican que la Red Bayesiana es superior a los otros dos modelos en todas las muestras de datos.

Bequé y Lessmann (2017) exploran el potencial de las Máquinas de Aprendizaje Extremo (ELM), que es un tipo de red neuronal artificial para la gestión del riesgo crediticio. Este nuevo método se compara con otras técnicas conocidas de credit scoring como son redes neuronales artificiales (ANN), k-vecinos más cercanos (KNN), árboles de decisión de regresión y clasificación (CART) y regresión logística regularizada (LR-R), para lo cual se usan tres conjuntos de datos reales. Los autores enfocan su estudio en determinar la facilidad de uso, la complejidad computacional y la precisión discriminativa de la predicción. En los experimentos se emplean dos medidas para las comparaciones: porcentaje correctamente clasificado y el área bajo la curva ROC. Se hacen comparaciones en forma individual y con clasificadores de conjunto, y los resultados muestran que no hay una clara superioridad entre los clasificadores individuales, pero el método propuesto tiene un rendimiento competitivo o superior a sus pares.

Tian, Yong y Luo (2018) proponen un nuevo enfoque basado en Maquina con Soporte Vectorial (SVM) donde la separación de las clases se hace mediante una superficie cuadrática difusa, prescindiendo del uso de los núcleos. Este modelo no solo se desempeña muy bien en la clasificación, también maneja algunos problemas importantes de los SVM clásicos, como buscar funciones del kernel adecuadas, además, elimina el mal efecto de los valores atípicos en la calificación de crédito. La principal contribución de esta investigación es desarrollar un nuevo enfoque para la inferencia de rechazo basado en el modelo Fuzzy Quadratic Surface Support Vector Machine

(FQSSVM) de núcleo libre. Este modelo en lugar de escoger un núcleo de varias alternativas utiliza directamente una superficie cuadrática para la separación. Los resultados muestran que el modelo FQSSVM logra una mejor clasificación, comparado con otros métodos para un mismo conjunto de datos. FQSSVM no solo tiene bajo error de tipo I, también tiene un error de tipo II mucho más bajo en todos los casos. Así, este nuevo método es más práctico en inferir si un solicitante debe ser aceptado o rechazado.

2.1.2. Clasificadores de conjunto

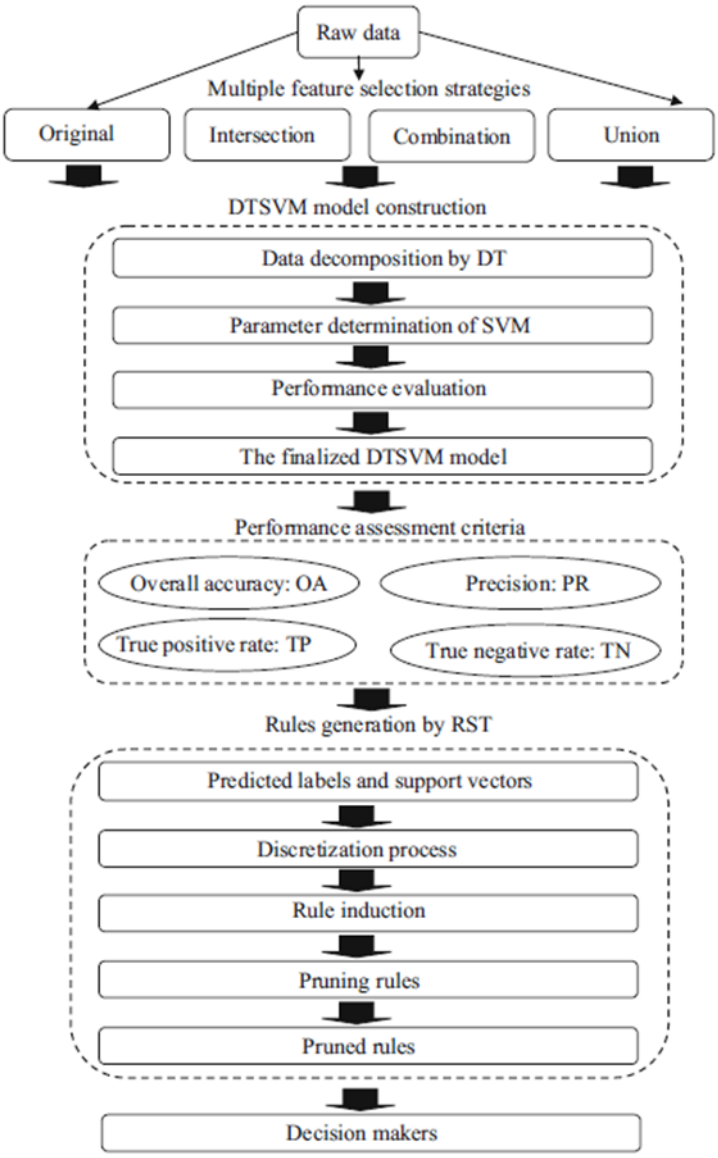
Wang, Hao, Ma y Jiang (2011) examinan el desempeño de los diferentes métodos de conjunto para el campo de la evaluación de crédito en términos de precisión media, error de tipo I y error de tipo II. Trabajan con 3 conjuntos de datos de crédito y se eligen 18 variables financieras para la puntuación de crédito. En los experimentos se eligen cuatro métodos de aprendizaje base: Regresión Logística, Árboles de decisión, Redes Neuronales y Máquina con Soporte Vectorial. Como resultado Bagging se comporta mejor que Boosting en los 3 conjuntos de datos, Stacking y Bagging DT tienen mejor desempeño en los indicadores de rendimiento precisión media, error tipo I y error tipo II. De los cuatro aprendices base, el Árbol de decisión tiene el mejor desempeño en los 3 indicadores de rendimiento.

Cubiles-De-La-Vega, Blanco-Oliver, Pino-Mejías y Lara-Rubio, J. (2013) construyen un conjunto amplio de modelos de calificación de crédito de las instituciones de microfinanzas dentro del marco de aprendizaje estadístico. Analizan un conjunto de datos de microcréditos de una microfinanciera peruana con información de clientes del período 2003-2008. Se construyen diversos modelos usando técnicas paramétricas como Análisis Discriminante Lineal, Análisis Discriminante Cuadrática y Regresión Logística, y técnicas no paramétricas como Redes Neuronales, Máquina con Soporte Vectorial, Bosques aleatorios, método Bagging y método Boosting. Para la evaluación de los modelos se usan el área bajo la curva ROC (AUC) y el costo esperado de clasificación errónea. Los resultados obtenidos muestran que los modelos no paramétricos tienen un mayor AUC y costos de clasificación errónea más bajos que los enfoques clásicos. La aplicación de la red neuronal ayuda a reducir las pérdidas de las

microfinancieras de manera significativa, por lo tanto, constituyen una ventaja competitiva sobre las otras microfinancieras que no aplican esta metodología. El modelo perceptron de tres capas, con 20 nodos en la capa de entrada, 3 nodos en la capa oculta y un nodo en la capa de salida, resulta ser el mejor modelo porque tiene el AUC más alto y menor costo esperado de clasificación errónea.

Pai, Tan y Hsu (2015) propusieron un modelo de calificación crediticia denominado M-DTSVM-RST, cuya estructura se muestra en la Figura 6.

Figura 6. **Estructura del modelo M-DTSVM-RST**



Fuente: Tomado de Pai et al.

Este modelo integra el Árbol de decisión SVM (DTSVM) para la tarea de selección de características (DTSVM) con la teoría de conjuntos ásperos para la generación de reglas para la toma de decisiones para el otorgamiento de los créditos.

Para probar la efectividad del modelo se usó un conjunto de datos reales y se comparó con otros modelos: Máquina con Soporte Vectorial multi-clase, Regresión Logística Multinomial, Análisis Discriminante Múltiple y Redes Neuronales Backpropagation. La evaluación del rendimiento se hizo con la exactitud general, la precisión, la tasa de verdaderos positivos y la tasa de verdaderos negativos. Los resultados experimentales muestran que el modelo M-DTSVM-RST propuesto supera a los otros enfoques de clasificación en todos los criterios de evaluación.

Ala´raj y Abbod (2016) afirman que los clasificadores de conjunto o sistemas de clasificadores múltiples han demostrado su capacidad para ser más precisos que los clasificadores individuales, por esta razón proponen un nuevo enfoque para la calificación crediticia basada en el consenso para combinar sistemas de clasificadores múltiples. Esta propuesta se basa en la combinación de un grupo de clasificadores diversificados, de modo que su combinación logre un mayor rendimiento que un clasificador individual, complementándose entre ellos. Para tal fin, emplean seis clasificadores básicos: Regresión Logística, Redes Neuronales, Máquinas con Soporte Vectorial, Bosques aleatorios, Árboles de decisión y Naïves Bayes. Para el entrenamiento y prueba de los clasificadores se emplean cinco conjuntos de datos reales. El rendimiento de los clasificadores se evalúa con la precisión promedio, el área bajo la curva (AUC), la medida H y el puntaje Brier (BS). Los resultados experimentales indican que ningún método de combinación tradicional logró mejores resultados que el mejor clasificador base, y que el enfoque de consenso ha demostrado ser un método de combinación confiable y eficiente al combinar clasificadores heterogéneos en varias medidas de evaluación.

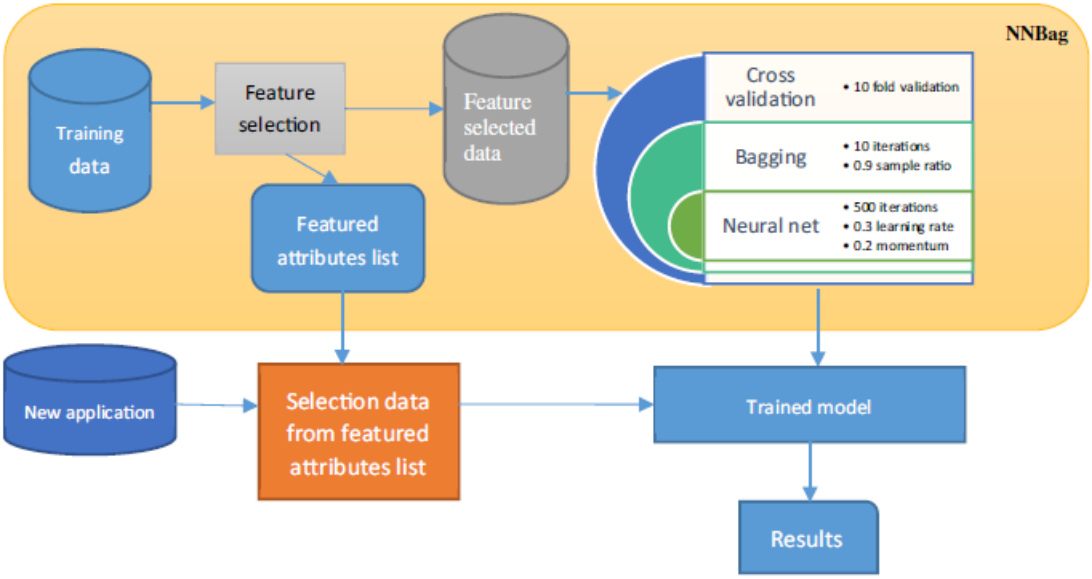
Feng, Xiao, Zhong, Qiu y Dong (2018) propusieron un nuevo método de clasificación de conjunto dinámico para la calificación crediticia basado en la probabilidad suave. Para generar los clasificadores de base utilizaron Árboles de decisión, Redes Neuronales y Máquina con Soporte Vectorial; además seleccionaron cuatro conjuntos homogéneos: BagDT, BagNN, BagSVM y Bosques aleatorios. Para los experimentos se usan diez conjuntos de datos de crédito reales, y el rendimiento de los clasificadores se

mide con la precisión general, el área bajo la curva ROC (AUC), la medida H, y el índice parcial Gini. El método propuesto se compara con 13 clasificadores de referencia, que incluyen clasificadores individuales, conjuntos homogéneos y conjuntos heterogéneos. Los resultados indican que el método propuesto es superior en la mayoría de las medidas de evaluación.

Zhang, He y Zhang (2018) desarrollan un modelo de conjunto denominado CF-GA-Ens que se basa en cinco clasificadores conocidos: Regresión Logística, Máquina con Soporte Vectorial, Red Neuronal, Árbol de decisión con gradiente boosting (GBDT) y Bosque aleatorio. Proponen un nuevo método que usa el algoritmo genético para seleccionar de múltiples algoritmos, el clasificador adecuado según las características del conjunto de datos. El modelo es probado con tres conjuntos de datos y para la evaluación se usan tres medidas de rendimiento: precisión, AUC (área bajo la curva ROC) y puntuación F. Los resultados muestran que cada algoritmo tiene una alta precisión predictiva en algunos de los conjuntos de datos, pero este clasificador conjunto tiene el mejor rendimiento en todos los conjuntos de datos y todas las métricas.

Dželihodžić et al. (2018) proponen un modelo que incluye la selección de características y un conjunto Bagging con Red Neuronal Backpropagation como clasificador base (NNBag). La Figura 7 muestra el proceso completo del modelo de evaluación crediticia. Este modelo es comparado con los clasificadores individuales: Árbol de decisión, Red neuronal y Regresión logística. Se emplean tres conjuntos de datos del mundo real para el experimento.

Figura 7. **Proceso del modelo Bagging con Red Neuronal.**



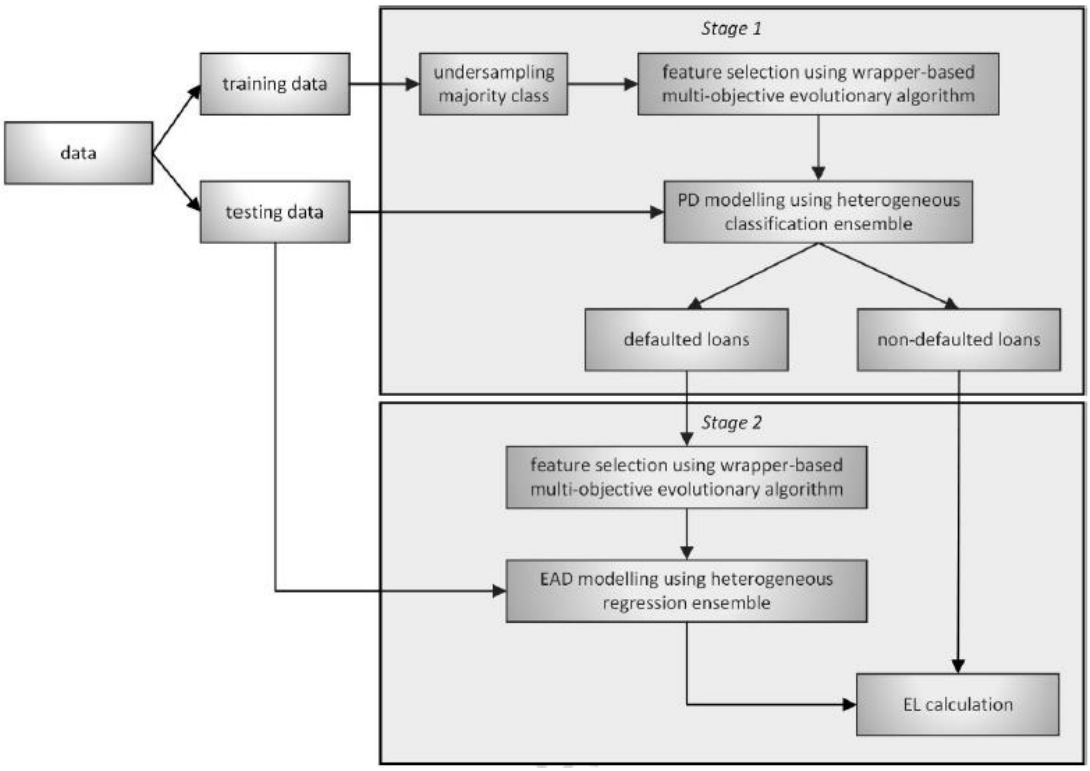
Fuente: Tomado de Dželihodžić et al. (2018)

Los resultados muestran que el modelo de conjunto propuesto tiene un buen desempeño en todos los criterios de evaluación, además se concluye que la selección de características mejora la evaluación de la solvencia.

Papouskova y Hajek (2019) afirman que para modelar el riesgo crediticio general de un préstamo de consumo en términos de la pérdida esperada (EL), se deben estimar tres parámetros clave del riesgo de crédito: probabilidad de incumplimiento (PD), pérdida dada el incumplimiento (LGD) y exposición al incumplimiento (EAD). Por lo tanto, ellos han propuesto un modelo de calificación del riesgo de crédito de dos etapas que integra (1) aprendizaje conjunto desequilibrado de clase para predecir PD, y (2) la predicción de EAD usando un conjunto de regresión. La Figura 8 muestra el marco conceptual del modelo propuesto. Para la parte experimental se usaron dos conjuntos de datos de crédito de consumo reales. En cuanto a las métricas, para evaluar la probabilidad de incumplimiento se emplearon la precisión de la predicción (Acc), el área bajo la curva ROC (AUC) y una métrica MC que combina LGD y el costo de oportunidad; para evaluar

la exposición al incumplimiento (EAD) y la pérdida esperada (EL) se usaron las métricas error absoluto medio (MAE), error cuadrático medio (RMSE) y el coeficiente de determinación. En el cálculo de PD, los resultados experimentales indican que el modelo de conjunto Stacking con RF superó a los otros clasificadores en todas las medidas de evaluación; para el cálculo de EAD, el modelo Stacking con LR funcionó mejor en un conjunto de datos, y el modelo Stacking con RF funcionó mejor en el otro conjunto de datos; y para el cálculo de EL, el método propuesto de dos etapas tuvo el mejor rendimiento en todas las medidas de evaluación.

Figura 8. **Marco conceptual para la modelización del riesgo de crédito.**



Fuente: Tomado de Papouskova et al. (2019)

Melo Junior, Maria Nardini, Renso, Trani y Antonio Macedo (2020) evalúan la combinación de métodos de selección dinámica, pre-procesamiento de datos y

conjuntos de generación de grupos, y proponen el método denominado Reduced Minority k-Nearest Neighbors (RMkNN). Ellos manifiestan que uno de los principales problemas en la calificación crediticia, es que manejan conjuntos de datos desequilibrados que generalmente contienen muchos préstamos al día y muy pocos préstamos no pagados, lo cual constituye un sesgo en los datos, por esta razón proponen esta técnica de selección dinámica para trabajar con conjuntos de datos desequilibrados.

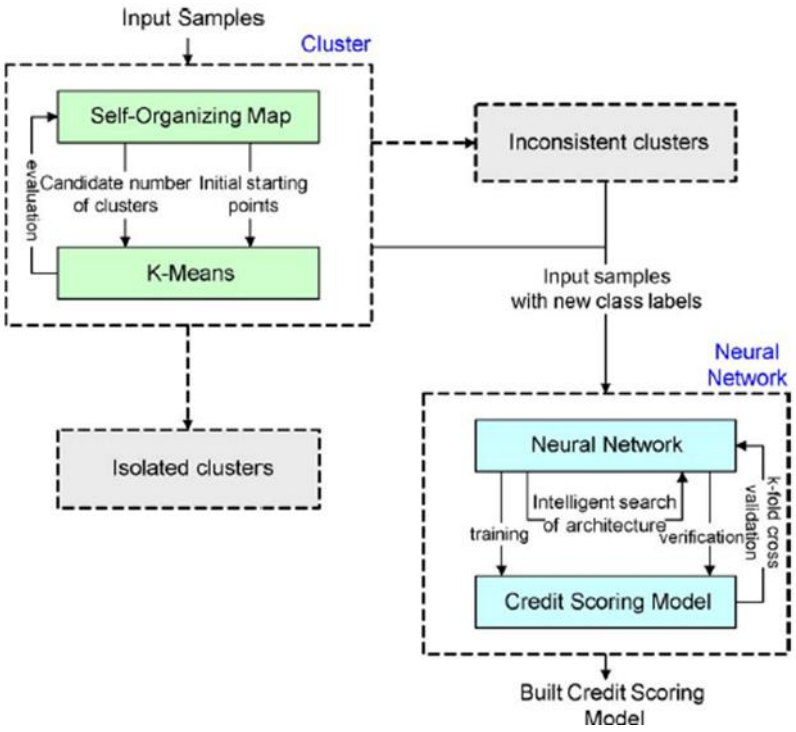
El algoritmo k-Nearest Neighbors (k-NN) se emplea para la selección dinámica de muestras, sin embargo, en conjuntos de datos desequilibrados, este algoritmo selecciona principalmente muestras de la clase mayoritaria, produciendo una pobre evaluación de la competencia de los clasificadores base, por tanto, para abordar el sesgo en los datos, desarrollan una modificación de este algoritmo proponiendo el k-NN de minoría reducida. Se experimenta con siete conjuntos de datos de crédito reales, y para la evaluación del rendimiento del modelo propuesto y de los otros comparativos, emplean el área bajo la curva ROC (AUC), la precisión equilibrada, medida H, media G, medida F y la recuperación del préstamo. Con respecto a los resultados obtenidos, al comparar k-NN de minoría reducida con k-NN en los siete conjuntos de datos, el primer algoritmo es más apropiado para manejar problemas de clasificación cuando la clase positiva mal clasificada es más alta. Finalmente se concluye que la combinación de k-NN de minoría reducida con técnicas de selección dinámica mejora el rendimiento de la predicción en todas las medidas de evaluación.

2.1.3. Clasificadores híbridos

Hsieh (2005) presenta un sistema de minería híbrida para el diseño de un modelo de credit scoring, para lo cual emplea la Red Neuronal y el algoritmo de agrupamiento K-means para pre-procesar las muestras en grupos homogéneos. En la Figura 9 se muestra el sistema propuesto. La propuesta del autor integra una técnica de agrupamiento y la red neuronal para la construcción del modelo. Para el entrenamiento y prueba del modelo se emplean dos conjuntos de datos reales. Los resultados obtenidos de las pruebas demuestran que el enfoque híbrido es simple pero eficiente para el *credit*

scoring; el uso de técnicas de agrupamiento permite identificar muestras no representativas, lo cual es valioso, porque permite la construcción de un modelo de red neuronal de alta precisión para la calificación de créditos.

Figura 9. **Sistema de minería híbrida para credit scoring.**



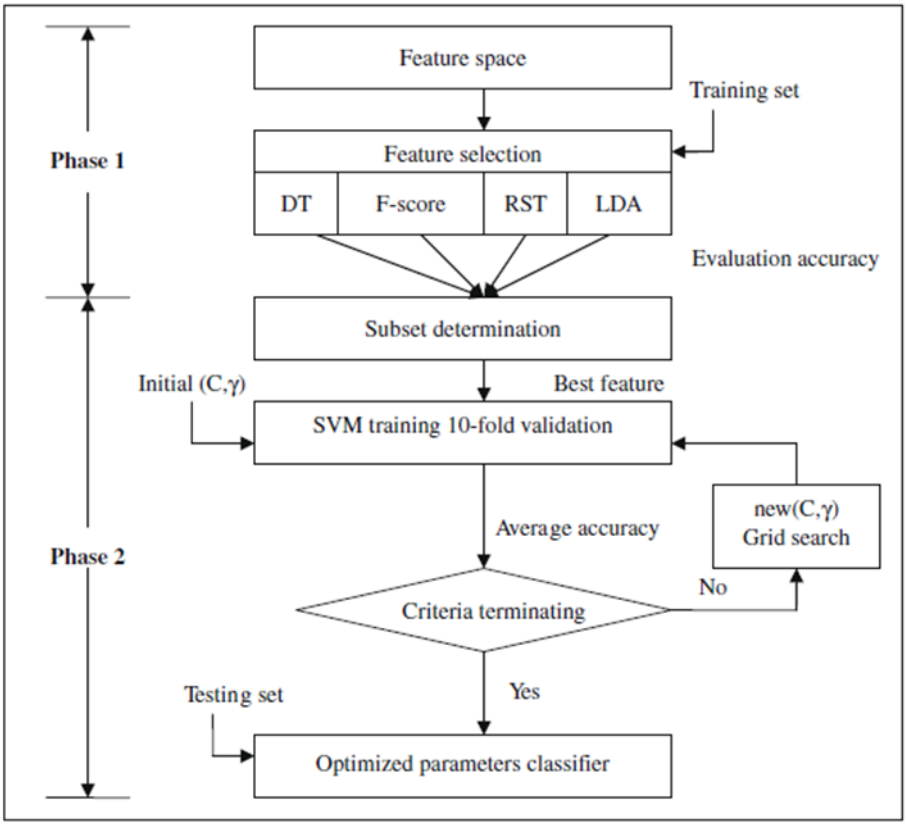
Fuente: Tomado de Hsieh (2005)

Finlay (2010) evalúa varios sistemas clasificadores múltiples para determinar el riesgo crediticio. Los resultados sugieren que algún clasificador múltiple tiene un rendimiento significativamente mejor que un clasificador solo. En su experimento usa dos conjuntos de datos y cinco métodos básicos para la construcción de los clasificadores: Regresión Logística, Análisis Discriminante Lineal, Clasificación y Árboles de regresión, Redes Neuronales y K-vecino más cercano. En general Bagging y Boosting superan a otros clasificadores múltiples y el nuevo algoritmo Error Trimmed Boosting

supera a Bagging y AdaBoost por un margen significativo, siendo AdaBoost el algoritmo boosting más conocido.

Chen y Li (2010) desarrollan un modelo de credit scoring empleando una arquitectura de clasificación híbrida de dos fases mediante cuatro enfoques que se combinan con el clasificador Máquina con Soporte Vectorial (MSV). Esta arquitectura se muestra en la Figura 10. El clasificador MSV se combina con el método estadístico Análisis Discriminante Lineal (ADL), Árbol de decisión (AD), Conjuntos rugosos y F-score para optimizar la selección de características y mejorar la precisión del modelo.

Figura 10. **Arquitectura del modelo propuesto basado en enfoques.**



Fuente: Tomado de Chen et al. (2010)

Se obtienen así los enfoques: "ADL + MSV", "AD+ MSV", "Conjuntos rugosos + MSV" y "F-score + MSV". Para los experimentos se emplean dos conjuntos de datos reales, se usa como métrica de evaluación la precisión y el área bajo la curva ROC (AUC). Los resultados obtenidos indican que el modelo ADL+MSV es ligeramente superior a los otros tres modelos. Una conclusión importante de este estudio es que, dado cualquier algoritmo de clasificación, se debe utilizar el método de selección de características apropiada para mejorar el rendimiento del modelo.

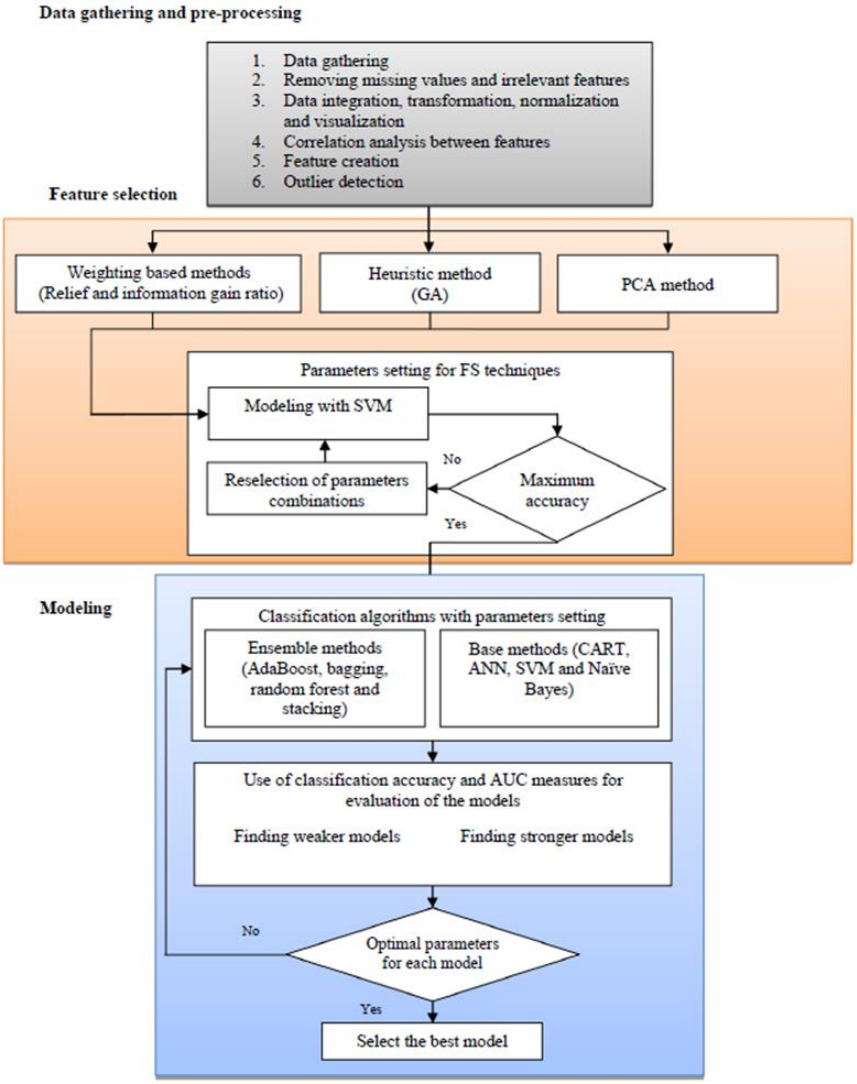
Oreski (2014) basándose en las técnicas híbridas de optimización combinatoria, propone una técnica híbrida de selección de características y la clasificación para la evaluación del riesgo crediticio, para lo cual hace la hibridación del Algoritmo genético con la Red Neuronal (AGH-RN). Para experimentar con el modelo propuesto se usa un conjunto de datos reales, y se usa la precisión para su evaluación. Los resultados obtenidos indican que el algoritmo AGH-NN es una alternativa aceptable para optimizar el subconjunto de características y los parámetros de redes neuronales para la evaluación del riesgo de crédito, proporcionando la mejor precisión media.

Oreski y Oreski (2014) proponen un Algoritmo genético híbrido con Redes Neuronales (AGH-RN) para identificar un subconjunto óptimo de características y aumentar la precisión en la evaluación del riesgo crediticio. Los resultados luego de los experimentos con dos conjuntos de datos reales indican que la técnica AGH-RN, en promedio muestra mejores resultados en términos de precisión que la técnica no híbrida.

Koutanaei, Sajedi y Khanbabaei (2015) desarrollaron un modelo híbrido para la selección de características y un clasificador de conjunto para credit scoring, la Figura 11 muestra este modelo. Para la selección de características emplea cuatro algoritmos, siendo el Análisis de componentes principales (PCA) el mejor. Para el clasificador conjunto usan cuatro algoritmos para métodos ensamblados: AdaBoost, Bagging, Bosques aleatorios y Stacking; y cuatro métodos básicos: Árbol de clasificación y regresión (CART), Redes neuronales artificiales (ANN), Máquina con soporte vectorial (SVM) y Naïve Bayes. El modelo es entrenado y probado con un conjunto de datos del Banco de desarrollo de exportaciones de Irán. Los resultados obtenidos indican que el

algoritmo PCA tiene el mejor rendimiento, porque las características seleccionadas conducen a una mayor precisión y medida AUC. ANN tiene la mejor precisión de clasificación que los métodos básicos. Además, ANN-AdaBoost tiene el mejor rendimiento que los otros algoritmos base y ensamblados.

Figura 11. **Diagrama de bloque del modelo propuesto.**

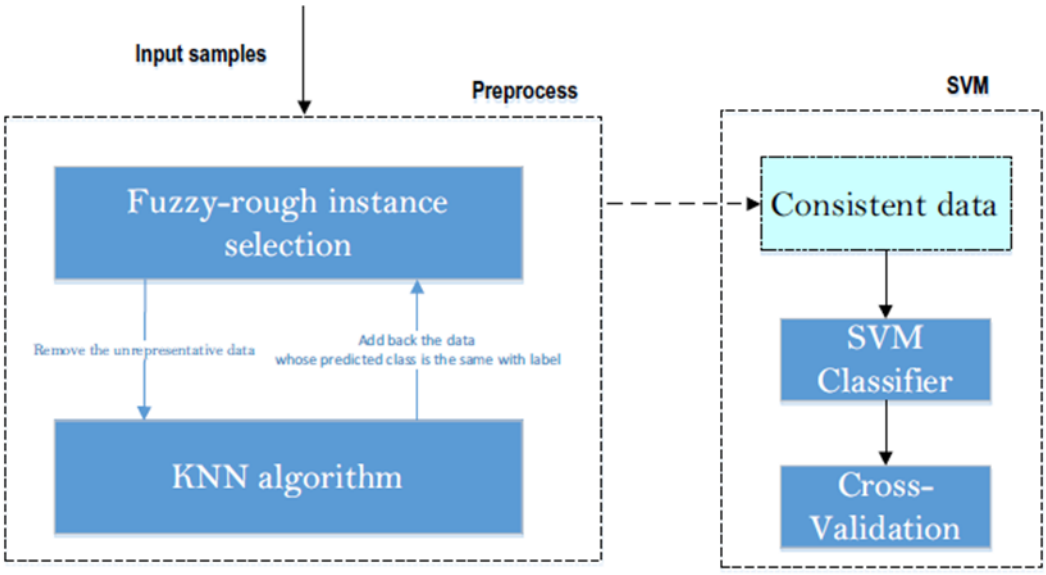


Fuente: Tomado de Koutanaei (2015)

AghaeiRad, Chen y Ribeiro (2016) desarrollan un modelo híbrido de calificación crediticia e ilustran cómo el aprendizaje no supervisado basado en un mapa auto-organizado (SOM) puede mejorar la capacidad discriminante de la Red Neuronal Feedforward (FNN). Los SOMs se utilizan para agrupar los datos y transferir conocimiento a la red FNN. Se usan cuatro conjuntos de datos reales para las pruebas. Las métricas para comparar el rendimiento son la exactitud, la sensibilidad y la especificidad. Variando los parámetros (tamaño de conjuntos de datos de entrenamiento, número de neuronas de FNN y SOM), los resultados demuestran que siempre el modelo de clasificación híbrida (FNN + SOM) propuesta produce un mejor rendimiento que el modelo FNN en todas las métricas.

Liu y Pan (2017) presentan un clasificador híbrido basado en un algoritmo de agrupación para identificar y procesar las instancias no representativas en agrupaciones aisladas e inconsistentes. Utilizan el mapa auto-organizado (SOM) para determinar el número de grupos y puntos de partida de cada grupo, luego usan el algoritmo k-means para generar grupos de instancias que pertenecen a nuevas clases y eliminar las instancias no representativas de cada clase, una vez que se tiene la data consistente, emplean la Máquina con Soporte Vectorial (SVM) con núcleo polinomial para hacer las predicciones, la arquitectura del algoritmo hibrido se muestra en la Figura 12. Para el experimento se usan dos conjuntos de datos reales, los resultados obtenidos se comparan con los clasificadores Análisis discriminante lineal, Regresión logística y Redes neuronales, y el clasificador híbrido propuesto tiene la mejor capacidad de calificación crediticia en términos de la tasa de clasificación general.

Figura 12. **Arquitectura del algoritmo hibrido.**

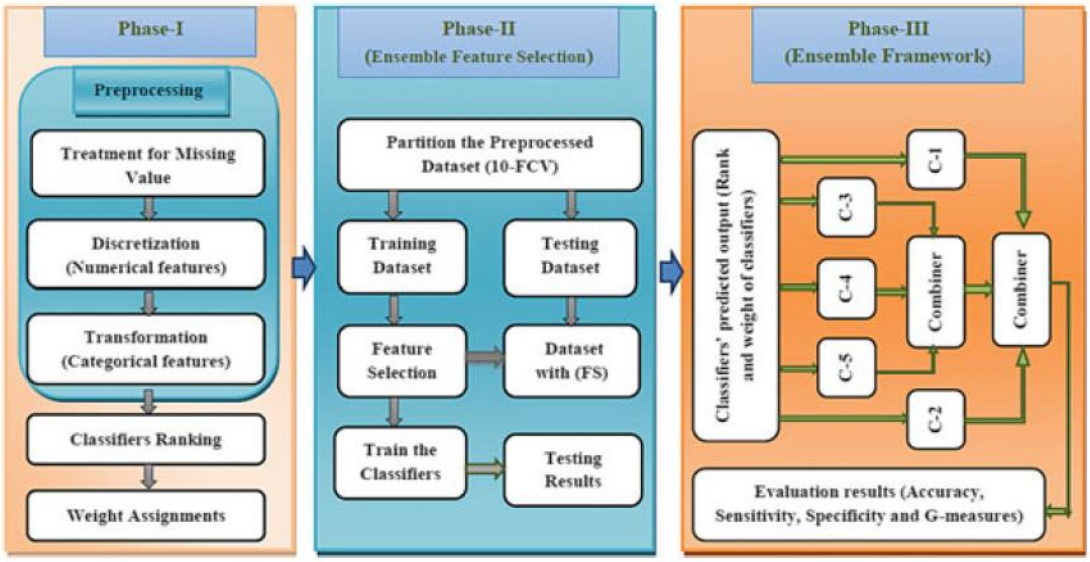


Fuente: Tomado de Liu et al.

Armaki, Fallah, Alborzi y Mohammadzadeh (2017) proponen un método híbrido para puntuación de crédito basado en la combinación de híbrido tradicional y métodos de apilamiento de conjuntos. El nuevo método híbrido es una mezcla de híbrido tradicional y el método de conjunto Stacking. En este proceso, en lugar de usar un solo algoritmo de clasificación en la primera parte del modelo híbrido, utilizan un método de conjunto Stacking (apilamiento). En la segunda parte, varias técnicas de agrupamiento se utilizarán indistintamente para encontrar qué combinación de algoritmos produce los mejores resultados. Se compara varias versiones del modelo híbrido propuesto seleccionando varias mezclas de algoritmos de clasificación y agrupamiento para encontrar el mejor modelo. El poder predictivo de los modelos desarrollados se mide a través de la tasa de precisión de predicción, medida F y errores tipo I y II. El modelo híbrido de meta-aprendizaje propuesto supera todos los clasificadores referenciados en la literatura en términos de tasa de precisión y errores tipo I/II. Este modelo también supera los mejores modelos utilizados en la literatura relevante en términos de tasa de precisión.

Tripathi, Edla, Cheruku y Kuppili (2019) desarrollan un modelo híbrido, combinando la selección de características y un marco clasificador multicapa de conjunto para mejorar el rendimiento predictivo de credit scoring. Este modelo se compone de tres fases tal como se muestra en la Figura 13.

Figura 13. **Marco híbrido propuesto para credit scoring**



Fuente: Toma de Tripathi et al.

Como medidas de rendimiento se emplean la exactitud, la sensibilidad, la especificidad y la medida G. Se consideran ocho clasificadores individuales: Análisis Discriminante Cuadrático, Naïve Bayes, Red Neuronal Feed forward Multicapa, Red Neuronal de tiempo retardado, Red Neuronal de tiempo retardado distribuido, Árboles de decisión, Máquina con Soporte Vectorial, y K-vecinos más cercanos, que luego son categorizados. El modelo se valida con cuatro conjuntos de datos reales, y luego de los experimentos se concluye que el modelo hibrido propuesto mejoró el rendimiento de los clasificadores individuales en términos de la precisión de la predicción.

2.2. MARCO TEÓRICO

2.2.1. Credit Scoring

El credit scoring también conocido como calificación crediticia, es una metodología de evaluación crediticia para medir la solvencia de un cliente, que consiste en asignar un puntaje al potencial deudor. El credit scoring se basa principalmente en el historial de préstamos del solicitante.

En la investigación de Rayo *et al.* (2010), se afirma que el credit scoring es una estimación, en el momento de solicitar un crédito, de cuál será el comportamiento crediticio hasta su vencimiento, atendiendo al riesgo del cliente. Se evalúa a través de un modelo predictivo de comportamiento de pago mediante una puntuación que mide el riesgo de un prestatario y/o de la operación. Los modelos de credit scoring en microfinanzas sirven de soporte a la decisión del analista de créditos, pero no la sustituye.

Blanco *et al.* (2013) afirma que las EFdeM, para aumentar su eficiencia en todo su proceso, necesitan minimizar sus costos y controlar el riesgo de crédito si quieren sobrevivir en el largo plazo. Una forma para que las EFdeM sean más eficientes con el fin de competir con los bancos comerciales, es la implementación de sistemas de credit scoring automáticos para evaluar sus solicitantes de crédito desde que la puntuación de crédito reduce el costo de análisis del crédito, mejora el flujo de caja, permite decisiones de crédito más rápido, reduce las pérdidas, y también resulta en el control más estricto de las cuentas existentes y la priorización de la colección de amortización. Según Bekhet et al. (2014), el modelo de credit scoring es un sistema de soporte de decisiones que ayuda a los gerentes en el proceso de toma de decisiones financieras. Con el rápido desarrollo de la industria crediticia, los modelos de calificación crediticia se utilizan en las decisiones relacionadas con la evaluación de admisión de créditos. Estos modelos se desarrollan para clasificar las solicitudes de crédito como "aceptadas" o "rechazadas" con respecto a las características de los solicitantes, como la edad, los ingresos, estado civil, etc. El objetivo del modelo de credit scoring es determinar la capacidad del

solicitante de crédito para pagar las obligaciones financieras mediante la evaluación del riesgo crediticio de la solicitud de préstamo.

Por su parte, Abid *et al.* (2016) define credit scoring como un conjunto de modelos de decisión con técnicas subyacente específicas que constituyen un sistema de soporte que ayuda a los prestamistas en el proceso de toma de decisiones financieras.

Para Zhang *et al*. (2018) el credit scoring es un modelo de toma de decisiones donde las solicitudes de préstamos de clientes "buenos" son aceptados y las solicitudes de préstamos de clientes "malos" son rechazadas mediante el uso de técnicas de análisis estadístico o algoritmos de máquinas de aprendizaje, que pueden reducir la posibilidad de préstamos morosos.

En conclusión, el credit scoring es un sistema que clasifica a los solicitantes de crédito en dos categorías: aquellos que tienen una alta probabilidad de cumplir sus obligaciones financieras están etiquetadas como “buenos” y aquellas que tienen una baja probabilidad de cumplir con estas obligaciones, se clasifican como “malos”.

2.2.2. Microfinanzas

Las microfinanzas son aquellos servicios y productos financieros orientados hacia el desarrollo de las pequeñas economías. Estas operaciones financieras de menor escala incluyen acceso a los microcréditos, cuentas de ahorros, transferencias de dineros, seguros y otros servicios que se brindan a las personas naturales y las microempresas. En el marco de la globalización, las microfinanzas se han constituido como un enfoque de las finanzas que apuntan a motivar la inclusión y la democratización de los servicios financieros para aquellos sectores generalmente excluidos por la banca comercial tradicional.

Dixon et al. (2006) afirma que las microfinanzas se basan en proporcionar pequeños préstamos, para los pobres y muy pobres para permitirles obtener ingresos adicionales invirtiendo en la fundación o crecimiento de "microempresas". En términos más generales, su objetivo es proporcionar microcréditos, ahorros, y otros servicios

financieros a los pobres. Opera bajo la premisa de que los pobres invertirán préstamos en microempresas, reembolsarán esos préstamos de las ganancias, y sus negocios crecerán. Estas expectativas se basan en la premisa que los pobres serán "empoderados", y animados a participar para auto gestionarse ocupaciones.

En las últimas décadas las microfinanzas han experimentado cambios en su estructura, de tal forma que se han alejado de los subsidios estatales o de las fuentes cooperantes, para acercarse a la autofinanciación que le permita la sostenibilidad en el tiempo, sin abandonar sus características básicas.

2.2.3. Microcrédito

Los microcréditos son pequeños préstamos dirigidos a personas naturales o jurídicas, que tienen dificultad para acceder a la banca comercial tradicional, y que son destinados al financiamiento de actividades en pequeña escala de producción, comercialización o servicios, cuya fuente principal de pago constituye el producto de las ventas o ingresos generados por dichas actividades. Estos préstamos generalmente se destinan a la financiación de pequeños proyectos de autoempleo que generan ingresos y posibilitan la autonomía económica de los beneficiarios y sus familias. Los microcréditos financian actividades económicas de pequeño tamaño, conocidas como “microempresas”, que se caracterizan por estar gestionadas por una persona o un grupo familiar, con reducido nivel de activos y escasa formación técnica gerencial, y que generalmente operan en la informalidad. La mayoría de los bancos comerciales no está dispuesto a otorgar créditos a las personas con escasos recursos, por el alto riesgo en la devolución del préstamo y por el reducido beneficio de trabajar con créditos a pequeña escala.

En América Latina la definición de microcrédito suele estar ligada a diferentes características de estos productos. En algunos países se utilizan los montos como base para definir un crédito como microcrédito, en otros es el sujeto, el destino del crédito, las actividades a financiar o la fuente de repago el aspecto que definirá un financiamiento como parte de este segmento.

Barboza *et al.* (2009) en su investigación sobre la Reducción de la pobreza mediante préstamos grupales afirma que Mohammed Yunus y el Grameen Bank en Bangladesh, fue el pionero en el desarrollo de los microcréditos. Este tipo de crédito fue creado para reducir la pobreza a través de la provisión de recursos financieros a los más pobres. A partir de esta iniciativa han surgido programas de microcrédito en muchos países y son una opción viable para las personas con habilidades empresariales y proyectos de negocios prometedores. Las organizaciones no gubernamentales gestionan programas de microcrédito y reciben apoyo financiero de donantes a tasas de interés preferenciales o cero, para luego prestar a las personas más pobres. La viabilidad financiera de los programas de microcrédito es crucial para su sostenibilidad, y esto se logra con altas tasas de reembolso. En América Latina, las instituciones de microcrédito más conocidas son BancoSol en Bolivia, FINCA en Costa Rica, ACCION International, y Compartamos en México. El objetivo que distingue a estas instituciones es proveer de recursos financieros para aquellos considerados "no financiables" por otros. Con respecto a los resultados, de acuerdo a la hoja informativa del evento de alto nivel de las Naciones Unidas, el 65% de todos los prestatarios del Grameen Bank “han logrado para salir de la pobreza extrema". Finalmente se señala que para la supervivencia de los programas de microcrédito y lograr el objetivo final de alivio a la pobreza, es fundamental mantener altas tasas de reembolso.

Por su parte Nanayakkara et al. (2015) afirman que el concepto de las microfinanzas se introdujo en el mundo en 1976 por el profesor Muhammad Yunus cuando creó el Grameen Bank en Bangladesh para ofrecer pequeños préstamos sin garantía a los pobres. Los préstamos ayudaron a los prestatarios para invertir en bienes para trabajar por cuenta propia y mejorar su nivel de vida, mientras pagaban el préstamo. Luego de tres décadas, la base de datos de clientes del banco había crecido a 6,9 millones de prestatarios y los desembolsos totales de préstamos ascendió a US $ 5.95 mil millones (Grameen Bank, 2007). Este concepto destruyó el mito de que los pobres no son dignos de crédito y Yunus fue galardonado con el Premio Nobel en 2006. El éxito fenomenal del Grameen Bank ha llevado a un rápido crecimiento de las instituciones microfinancieras (IMF) en todo el mundo.

En la actualidad los microcréditos han traspasado las fronteras de los países pobres hacia los países desarrollados, éste hecho demuestra el éxito que han tenido estos créditos a nivel global.

En el Perú, según Portocarrero, Trivelli y Alvarado, J. (2002), las instituciones de microfinanzas (IMF) fundamentalmente se agrupan en dos. Un grupo está constituido por las entidades bancarias y financieras privadas que no tienen mayores requisitos de capital mínimo, y que se han especializado en atender a los sectores de bajos ingresos, como Mibanco y Financiera Solución. El otro grupo lo conforman los intermediarios financieros no bancarios (IFINB) como las Cajas Rurales de Ahorro y Crédito (CRACs), las Cajas Municipales de Ahorro y Crédito (CMACs) y las Entidades de Desarrollo de la Pequeña y Microempresa (EDPYMES), que tienen exigencias de capital mínimo más reducido, operan a escala regional y están facultadas para realizar un conjunto más limitado de operaciones.

Se estima que el Banco del Trabajo llegó a contar con 35 mil clientes en el crédito PYME, con una cartera vigente de US$ 20 millones a fines de 1999 y un financiamiento promedio de US$ 571 por cliente. Así, se orienta claramente a los sectores de las microempresas de menor dimensión, como el resto de los intermediarios especializados en la banca de consumo. También se aprecia que las colocaciones en dólares de Mibanco a las PYMES, las CMACs y las EDPYMES se han incrementado en 1999 en un 24.6 % en dólares, a pesar del contexto recesivo imperante. Se advierten algunas diferencias interesantes en la composición de la clientela de las diferentes entidades especializadas. Mibanco otorga los créditos más pequeños, por lo que muestra los mayores niveles de cobertura. A su vez, las CRACs registran los mayores promedios, lo que indicaría que se dirigen a sectores más consolidados de las PYMES. El resto de estas entidades se sitúa cerca del promedio.

Tabla 3. Ranking de créditos a Microempresas por Empresa Financiera

Empresa Financiera	Créditos (miles de soles)		Variación	Participación del mercado
	Ago 2020	Jul 2020		
Compartamos	*1,134,170*	*1,158,385*	*-2.09%*	*44.85%*
Confianza	*641,114*	*629,694*	*1.81%*	*25.35%*
Proempresa	*206,399*	*175,871*	*17.36%*	*8.16%*
Credinka	*194,302*	*195,615*	*-0.67%*	*7.68%*
Crediscotia	*176,421*	*191,330*	*-7.76%*	*6.98%*
Qapaq	*80,542*	*69,746*	*15.48%*	*3.18%*
Efectiva	*79,964*	*78,926*	*1.31%*	*3.16%*
MAF	*16,112*	*14,434*	*11.63%*	*0.64%*
Oh!	*0*	*0*		*0.00%*
Amérika	*0*	*0*		*0.00%*
Total	*2,529,024*	*2,514,001*	*0.60%*	*100.00%*

Fuente: SBS Elaborado por: ICC Instituto de Créditos y Cobranzas. Extraído de Microfinanzas.pe (Micro-finanzas en el Perú 2020)

En general, el monto de los microcréditos colocados por las entidades financieras muestra una tendencia creciente en los últimos años, en la Tabla 3 se muestra la participación en el mercado que tienen las diferentes entidades financieras, y en la Tabla 4 se muestra los créditos otorgados por las diferentes instituciones de microfinanzas, haciendo una comparación al mes de abril de los años 2019 y 2020.

Tabla 4. Créditos directos del Sistema Financiero al 30 de abril del 2020

Institución	Monto (miles de soles)		Crecimiento
	Abr 2019	Abr 2020	
Cajas Municipales	21,896,985	22,837,996	4.30%
Cajas Rurales	2,374,091	2,282,789	-3.85%
Edpymes	2,307,607	2,554,474	10.70%
Financieras	13,189,913	13,627,628	3.32%
Financiera esp MF	10,339,757	10,456,568	1.13%
Financiera no esp MF	2,850,156	3,171,060	11.26%
Mibanco	10,132,422	10,828,700	6.87%
Total IMF	47,050,861	48,960,527	4.06%
Banca	269,890,892	295,515,572	9.49%
Total Sistema Financiero	309,659,487	336,818,458	8.77%

Fuente: SBS, de Asomif (Asociación de Instituciones de Micro-finanzas del Perú. En: http://www.asomifperu.com/web/

2.2.4. Morosidad

La morosidad se define como una situación en la que un deudor no cumple con el pago al vencimiento de una obligación. El término moroso se utiliza indistintamente para referirse a conceptos diferentes. Por un lado, el retraso en el cumplimiento de una obligación se denomina jurídicamente mora, y por consiguiente se considera moroso al deudor que se demora en su obligación de pago. Consecuentemente se considera que el cliente se halla en mora cuando su obligación está vencida y retrasa su cumplimiento de forma culpable. La mora del deudor en sí, desde el punto de vista formal, no supone un incumplimiento definitivo de la obligación de pago, sino simplemente un cumplimiento tardío de la obligación.

Rayo et al. (2010) consideran que el incumplimiento en el pago debe definirse con cautela, por lo que es necesario identificar todo atraso que conlleve un costo para el prestamista. Este incremento en el costo suele darse en términos de costos administrativos, debido a la necesidad de realizar un seguimiento y gestionar el pago de un crédito cuyo reembolso mantiene un retraso considerable. Por esta razón, definen el

atraso en el pago como un costo añadido para la micro-financiera con un mínimo de 30 días desde el vencimiento de al menos una cuota de amortización del microcrédito concedido.

2.2.5. Riesgo crediticio

El riesgo crediticio se define como la pérdida potencial ocasionada por el hecho de que un deudor incumpla con sus obligaciones de acuerdo con los términos establecidos en el contrato de préstamo. En toda operación de otorgamiento de un crédito está inherente el riesgo crediticio, es decir, el riesgo es una parte integrante de los servicios financieros.

Según Derelioglu et al. (2011) el riesgo de crédito es un término general que implica pérdidas futuras. El análisis del riesgo crediticio tiene como objetivo reducir las pérdidas futuras mediante la estimación del riesgo potencial y la denegación de la propuesta de crédito si el riesgo es más alto que un valor de tolerancia definido.

Moradi et al. (2019) define el riesgo crediticio como la probabilidad de impago o retraso en el pago por parte de los clientes o su incapacidad para pagar un préstamo. Según el acuerdo de Basilea 2, el riesgo de crédito es uno de los riesgos que enfrentan los bancos al asignar recursos, y cada banco necesita organizar y desarrollar su propio sistema interno de calificación crediticia con el que puedan analizar el riesgo crediticio.

2.3. MÁQUINA CON SOPORTE VECTORIAL

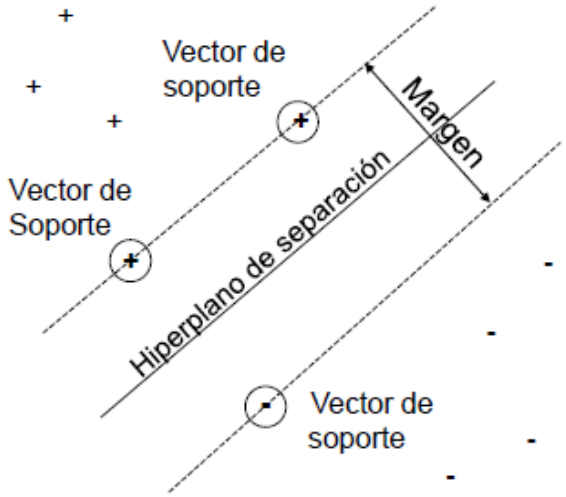
Las máquinas con soporte vectorial son métodos desarrollados en la década de los 90, originalmente como método de clasificación binaria y actualmente su uso se ha extendido a problemas de clasificación múltiple y regresión. Es un buen clasificador y es considerado uno de los referentes de Machine Learning. Este método se fundamenta en el máximo margen clasificador, que se basa a su vez en el concepto de hiperplano.

En clasificación las máquinas con soporte vectorial (MSV) se usan cuando se tienen datos exactamente de dos clases. El problema es separar las dos clases $A+ = \{ai$

/ yi = 1} y $A-$ = {ai /yi = −1}, mediante dos hiperplanos de separación óptima, que significa conseguir una franja o región de separación de las dos clases, la más amplia posible. La frontera de esta franja está determinada por vectores soporte de cada clase (Espinoza, 2020).

Una MSV clasifica los datos encontrando el mejor hiperplano que separa todos los puntos de datos de una clase de los de la otra clase. El mejor hiperplano para un MSV significa el que tiene el mayor margen entre las dos clases. Este margen es el ancho máximo de la franja paralela al hiperplano que no tiene puntos de datos interiores. Los vectores de soporte son los puntos de datos que están más cerca del hiperplano de separación; estos puntos están en la frontera de la franja. La Figura 14 ilustra estas definiciones, indicando con + los puntos de datos que están en la clase A+ y con – los puntos de datos que están en la clase A-.

Figura 14. **Hiperplano de separación entre las dos clases de datos.**



Fuente: Elaboración propia

Una recta o hiperplano ***L*** definida por

$$< c, x > + b = 0$$

determina tres regiones disjuntas en el espacio R^n, de tal manera que:

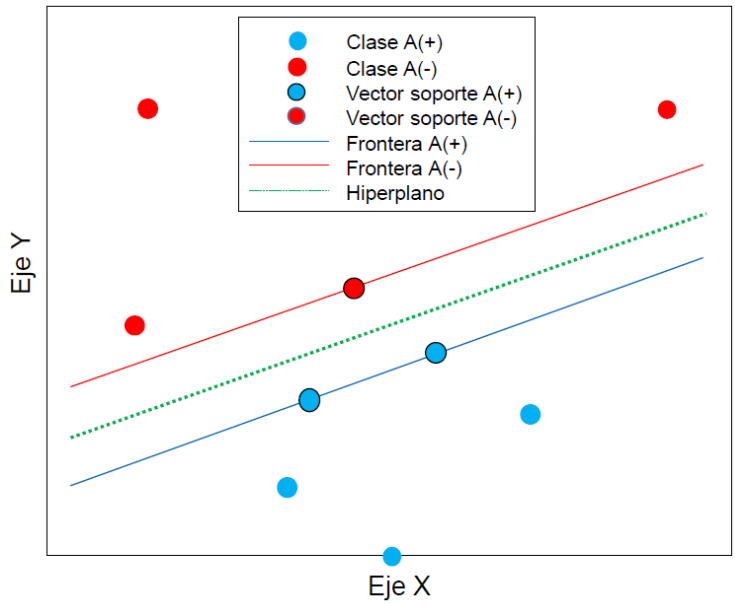
$< c, x > + b = 0$; define los puntos u que están en el hiperplano ***L***.

$< c, x > + b > 0$; define los puntos u que están en la clase A^+.

$< c, x > + b < 0$; define los puntos u que están en la clase A^-.

La figura 15 ilustra las clases, los vectores de soporte, los hiperplanos de frontera y el hiperplano separador en la máquina con soporte vectorial.

Figura 15. **Clases, vectores soporte e hiperplanos en la Máquina con Soporte Vectorial.**



Fuente: Elaboración propia

Con respecto al uso de la Máquina con Soporte Vectorial, Carmona (2014), afirma que, aunque originalmente las MSVs fueron pensadas para resolver problemas

de clasificación binaria, actualmente se utilizan para resolver otros tipos de problemas como regresión, agrupamiento, multiclasificación, etc. Las MSVs pertenecen a la categoría de los clasificadores lineales, puesto que inducen separadores lineales o hiperplanos, ya sea en el espacio original de los ejemplos de entrada, si éstos son separables o cuasi-separables (ruido), o en un espacio transformado (espacio de características), si los ejemplos no son separables linealmente en el espacio original. Mientras la mayoría de los métodos de aprendizaje se centran en minimizar los errores cometidos por el modelo generado a partir de los ejemplos de entrenamiento (error empírico), el sesgo inductivo asociado a las MSVs radica en la minimización del denominado riesgo estructural. La idea es seleccionar un hiperplano de separación que equidista de los ejemplos más cercanos de cada clase para, de esta forma, conseguir lo que se denomina un margen máximo a cada lado del hiperplano. Además, a la hora de definir el hiperplano, sólo se consideran los ejemplos de entrenamiento de cada clase que caen justo en la frontera de dichos márgenes. Estos ejemplos reciben el nombre de vectores soporte. Desde un punto de vista práctico, el hiperplano separador de margen máximo ha demostrado tener una buena capacidad de generalización, evitando en gran medida el problema del sobreajuste a los ejemplos de entrenamiento.

Desde un punto de vista algorítmico, el problema de optimización del margen geométrico representa un problema de optimización cuadrático con restricciones lineales que puede ser resuelto mediante técnicas estándar de programación cuadrática. La propiedad de convexidad exigida para su resolución garantiza una solución única, en contraste con la no unicidad de la solución producida por una red neuronal artificial entrenada con un mismo conjunto de ejemplos.

Según Dereliog et al. (2011), la MSV es un método basado en discriminante que intenta encontrar el hiperplano óptimo que maximiza la distancia entre los puntos de datos de diferentes clases en la clasificación. La distancia del hiperplano en cada lado se llama margen y el objetivo es maximizar el margen. El resultado de la clasificación de MSV es determinado por la salida, si es mayor que 0, entonces pertenece a la clase es 1 indicando que es un mal cliente, en otro caso, una salida igual a 0 indica que pertenece

a la clase 2, tratándose de un buen cliente. MSV también puede manejar problemas no lineales mapeando el espacio de entrada en un espacio no lineal por transformación no lineal. Diferentes núcleos tales como el polinomial o el núcleo de base radial es ampliamente utilizado para la transformación. Utilizamos SVM para fines de clasificación del riesgo de crédito con diferentes núcleos y comparar los resultados entre sí, también con otros clasificadores.

Para Han et al. (2013), la idea principal de la máquina de vector de soporte es minimizar el límite superior de la generalización del error, no del error empírico. Sin pérdida de generalidad, en un espacio bidimensional, si estas muestras de puntuación son lineales, escalables, el límite superior puede ser construido por (wx) + b = 1 y (wx) + b = –1, por lo que una función de decisión puede ser creada para especificar cuándo la aplicación del fragmento pertenece a cualquiera de los dos, I o J. Su definición es la siguiente: f(x) = sign ((wx) + b). Mientras que el vector w define el límite, para obtener dos límites superiores separados tan lejos como sea posible, el hiperplano óptimo puede obtenerse como una solución al problema de optimización:

$$\begin{aligned} &\max \tfrac{2}{\|w\|^2} \\ &s.t. \\ &y_i\big((w\cdot x_i)+b\big)\geq 1 \quad i=1,2,\cdots n \end{aligned}$$

El cual puede ser escrito como:

$$\begin{aligned} &\min \tfrac{1}{2}\|w\|^2 \\ &s.t. \\ &y_i\big((w\cdot x_i)+b\big)\geq 1 \quad i=1,2,\cdots n \end{aligned}$$

Entonces, un hiperplano de separación óptimo es uno que separa los datos con el margen máximo, es construido resolviendo un problema de optimización cuadrática con restricciones lineales, cuya la solución tiene una expansión en términos de un subconjunto de patrones de entrenamiento que se encuentran cerca del límite, y este subconjunto de patrones son llamados vectores soporte (SV).

2.4. REDES NEURONALES ARTIFICIALES

2.4.1. Neurona Artificial

Una neurona artificial se asemeja a una neurona bilógica en cuanto a sus componentes y al proceso de transferencia de información que ocurre dentro de la neurona.

Los componentes básicos de una neurona artificial son tres:

a) Un vector o matriz fila $w = [w_1, w_2, \ldots, w_n]$, denominada matriz de pesos.

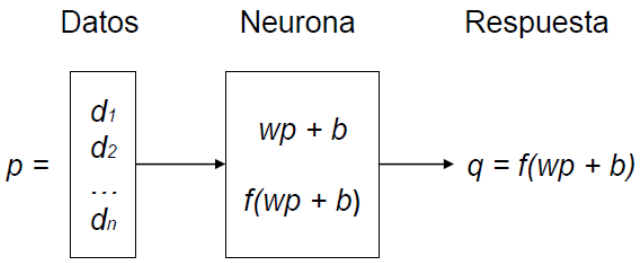
b) Un parámetro ***b*** denominado peso o sesgo de la neurona.

c) Una función real ***f*** llamada función de transferencia de la neurona.

La entrada o datos de ingreso de la neurona, es un vector o matriz columna $p = [d_1, d_2, \ldots, d_n]$. La respuesta o salida de la neurona es el número ***q***.

En una neurona artificial la entrada ***p*** se transmite a través de una conexión que multiplica su fuerza por el peso ***w***, para formar el producto ***wp***. Además, la neurona tiene un sesgo ***b***, que se adiciona al producto ***wp*** y esta suma constituye el argumento de la función de transferencia ***f***, que produce la salida ***q***, tal como se muestra en la Figura 16.

Figura 16. **Neurona simple**



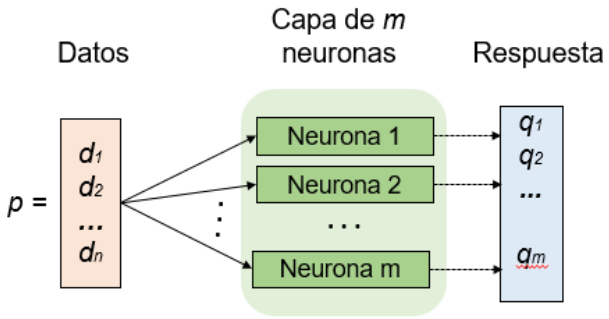
Fuente: Elaboración propia

La entrada neta de la función de transferencia es la suma de los valores ponderados de entrada wp y el sesgo b. Esta suma es el argumento de la función de transferencia f. Los parámetros w y b son ambos escalares ajustables de la neurona. La idea central de las redes neuronales es que dichos parámetros se pueden ajustar para que la red exhiba algún comportamiento interesante. Por lo tanto, se puede capacitar la red para que realice un trabajo en particular ajustando los parámetros de peso o sesgo, o tal vez la propia red ajustará estos parámetros para lograr algún fin deseado.

Capas de las neuronas artificiales

Una capa de neuronas está formada por *m* neuronas, dispuestas en paralelo, que operan independientemente, como se muestra en la Figura 17. En esta red, cada elemento del vector de entrada *p* está conectado a cada neurona. Es común que el número de elementos del vector *p* sea diferente del número de neurona. A su vez, cada neurona *j* tiene asociada una matriz de pesos Wj = [wj,1, wj,2, …, wj,n] más un sesgo *bj* y una función de transferencia *fj*.

Figura 17. **Capa de neuronas artificiales**



Fuente: Elaboración propia

La matriz $W = [w_{j,k}]$ es la matriz de pesos de la capa de neuronas, y está formada por las matrices filas de las *m* neuronas.

La matriz columna *b*, son los sesgos de la capa de neuronas.

La acción de la capa de neuronas sobre el vector ***p*** en una primera instancia es *Wp+b*. En una segunda instancia, cada componente de *Wp+b* es transformada por la función de transferencia f_j de la neurona, así la acción de la capa de neuronas sobre el vector *p* será:

$$q = f(Wp+b) = [\, f_1(w_1p+b_1),\ f_2(w_2p+b_2),\ \ldots,\ f_m(w_mp+b_m)]$$

q es un vector, y es la respuesta de la capa de neuronas para un vector de entrada ***p***.

Redes neuronales multicapas

Una red neuronal artificial (RNA) puede tener varias capas, cada capa tiene una matriz de pesos *W*, un vector de sesgos *b*, y un vector de salida *q*, por lo que los elementos de procesamiento en la red se encuentran agrupados por capas, tal como se indica en la Figura 18. De acuerdo a la ubicación de la capa en la RNA, éstas pueden denominarse así:

(i) *Capa de entrada:* Es la primera capa de neuronas. Esta recibe los datos de entrada a la red.

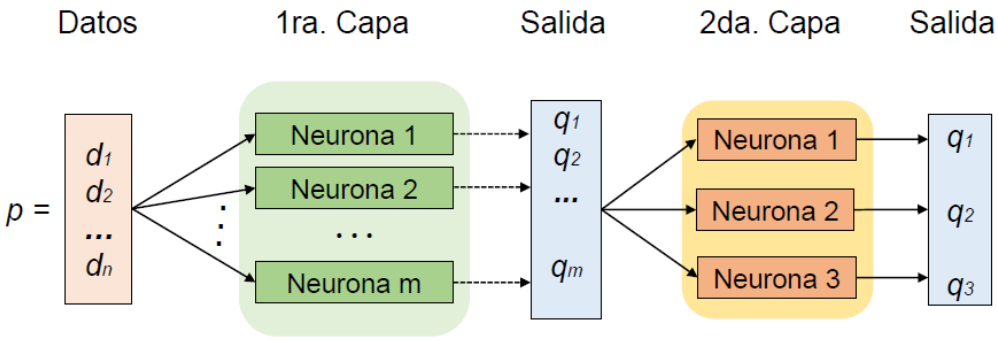
(ii) *Capas ocultas: Estas son las capas que siguen a la capa de entrada. No emiten las señales finales de la RNA.*

(iii) *Capa de salida: Es la última capa que sigue a las capas anteriores y es la que envía la respuesta o salida final de la RNA.*

Es común que diferentes capas tengan diferentes números de neuronas. Las capas de una red multicapa desempeñan funciones diferentes. Una capa que produce la salida de la red se denomina capa de salida. Todas las demás capas, a excepción de la capa de entrada, se llaman capas ocultas.

Finalmente, una red neuronal artificial es una concatenación de capas de neuronas.

Figura 18. **Red neuronal multicapa**



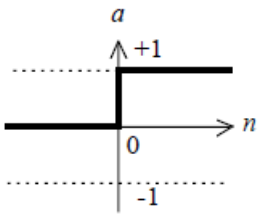
Fuente: Elaboración propia

Funciones de transferencia

Existen muchas funciones de transferencia empleadas en el diseño de redes neuronales, las funciones más utilizadas son:

a) Función de transferencia Limitador fuerte (hardlim)

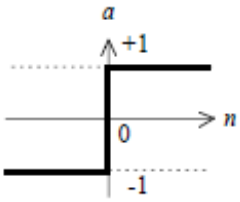
$$f(s) = \begin{cases} 0 \quad si \quad s<0 \\ 1 \quad si \quad s>=0 \end{cases}$$



Función de transferencia Limitador fuerte

b) *Función de transferencia Limitador fuerte simétrico (harlims)*

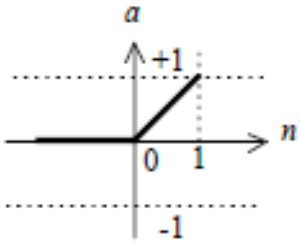
$$f(s) = \begin{cases} -1 \quad si \quad s<0 \\ \\ 1 \quad si \quad s>= 0 \end{cases}$$



Función de transferencia Limitador fuerte simétrico

c) *Función de transferencia Lineal positiva (poslin)*

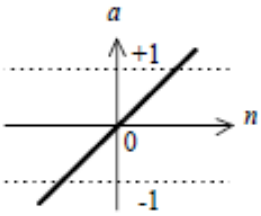
$$f(s) = \begin{cases} 0 \quad si \quad s<0 \\ \\ s \quad si \quad s>= 0 \end{cases}$$



Función de transferencia lineal positiva

d) *Función de transferencia Lineal (purelin)*

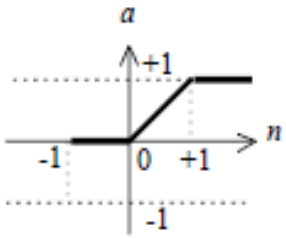
$$f(s) = s$$



Función de transferencia lineal

e) *Función de transferencia Lineal saturada (satlin)*

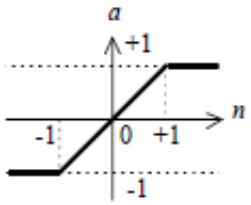
$$f(s) = \begin{cases} 0 & si \quad s<0 \\ s & si \quad 0<=s<=1 \\ 1 & si \quad 1<s \end{cases}$$



Función de transferencia lineal saturada

f) *Función de transferencia Lineal saturada simétrico (satlins)*

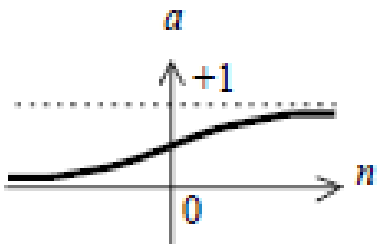
$$f(s) = \begin{cases} -1 & si \quad s<0 \\ s & si \quad 0<=s<=1 \\ 1 & si \quad 1<s \end{cases}$$



Función de transferencia lineal saturada simétrico

g) *Función de transferencia Sigmoidal (logsig)*

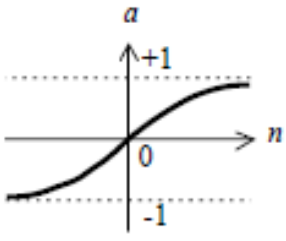
$$f(s) = \frac{1}{1+e^{-s}}$$



Función de transferencia sigmoidal

h) *Función de transferencia Tangente sigmoidal hiperbólica (tansig)*

$$f(s) = \tanh(s) \ \frac{e^{s}-e^{-s}}{e^{s}+e^{-s}}$$



Función de transferencia tangente sigmoidal hiperbólica

i) Función de transferencia Competitiva (compet)

$$f(s) = \begin{cases} 1 & \text{neurona con s máximo} \\ 0 & \text{en los demás casos} \end{cases}$$

2.4.2. Clasificación de las RNA según el tipo de aprendizaje

La regla de aprendizaje se define como el procedimiento para modificar los pesos y los sesgos de una red, por lo que este procedimiento también se denomina algoritmo de aprendizaje. El aprendizaje se aplica para entrenar la red para realizar alguna tarea. Las reglas de aprendizaje se dividen en dos categorías: aprendizaje supervisado y aprendizaje no supervisado. De aquí proviene la clasificación de las RNA que se presenta a continuación.

a) Redes Neuronales Supervisadas

Estas redes neuronales tienen una secuencia de vectores $T = [T_1, T_2, \ldots, T_M]$ que se denomina valor esperado de la red, donde T_k es el resultado de algún proceso realizado con la columna p_k de una matriz de datos $P = [p_1, p_2, \ldots, p_M]$, para cada $k = 1, 2, \ldots, M$.

De otro lado, para cada la columna p_k de P, una RNA devuelve un vector $q_k(x)$ donde x es un vector que está formado por los pesos y sesgos de todas las neuronas de la red. De este modo la respuesta de la red para la matriz de datos P que ingresa a ella, es otra matriz de vectores $Q(x) = [q_1(x), q_2(x), \ldots, q_M(x)]$.

El entrenamiento de la red neuronal supervisada consiste en presentar un vector de entrada a la red, calcular la salida de la red, compararla con el valor esperado, y el error o diferencia resultante se utiliza para ajustar los pesos y sesgos de la red de acuerdo con un algoritmo que tiende a minimizar el error.

El objetivo en el entrenamiento de la red neuronal supervisada es, por lo tanto, conseguir el mínimo global de la función de error E(x). Se trata del error en media cuadrática entre *T* y la respuesta de la red *Q*. Lo ideal sería conseguir un x_0 tal que $E(x_0) = 0$, pero esto no siempre es posible, entonces lo que se hace es minimizar la función E(x), buscando un x_0 de tal manera que $||E(x_0)|| \approx 0$. Esto implicaría que la respuesta de red está muy próxima al valor esperado. En este caso se dice que la RNA ha sido entrenada adecuadamente, en caso contrario no se habrá conseguido entrenar la red.

b) Redes Neuronales no Supervisadas

Las RNA de aprendizaje no supervisadas, son las que no requieren del vector de salida de valores esperados *T,* solo requieren de una matriz de datos *P*, y por lo tanto no se realizan comparaciones entre las salidas reales y las salidas esperadas. Estas redes tienen una sola capa, cuyo número de neuronas lo elige el usuario. En el entrenamiento, el algoritmo modifica los pesos de la red de forma tal que produzca vectores de salida consistentes. En el proceso de entrenamiento, se extraen las propiedades estadísticas del conjunto de vectores de entrenamiento y se realizan operaciones de agrupamiento de los vectores en grupos o clúster, clasificándose de esta manera los vectores de entrada en un número finito de clases.

2.4.3. Perceptrones

Esta red neuronal consta de una sola capa de neuronas conectado al vector de entrada *P* a través de la matriz de pesos *W*. Un perceptron es entrenado con un aprendizaje supervisado.

Las redes de perceptrones deben capacitarse con *adapt*, que presenta la entrada de vectores a la red uno a la vez y hace correcciones a la red basado en los resultados de cada entrada. El uso de *adapt* de esta manera garantiza que cualquier problema linealmente separable se resuelve en un número finito de presentaciones de entrenamiento. Los perceptrones también se pueden entrenar con la función *train*, al usar esta función, las entradas a la red se presentan en lotes, y hace correcciones a la

red basada en la suma de todas las correcciones individuales. Desafortunadamente, no hay pruebas de que dicho algoritmo de entrenamiento converja para los perceptrones. Por esta razón, no se recomienda el uso de *train* para perceptrones.

Las redes de perceptrones tienen limitaciones. Primero, el valor de salida de un perceptrón puede tomar solo uno de dos valores (0 o 1) debido al límite estricto de la función de transferencia. En segundo lugar, los perceptrones solo pueden clasificar conjuntos linealmente separables de vectores. Si se puede dibujar una línea recta o un plano para separar la entrada de vectores en sus categorías correctas, los vectores de entrada son linealmente separables. Si los vectores no son linealmente separables, el aprendizaje nunca llegará a un punto en el que todos los vectores están clasificados correctamente. Se debe tener en cuenta, sin embargo, que se ha demostrado que, si los vectores son linealmente separables, los perceptrones entrenados adaptativamente siempre encontraran una solución en un tiempo finito.

Los perceptrones de una sola capa pueden resolver problemas solo cuando los datos son linealmente separables. Una solución a esta dificultad es utilizar un método de pre-procesamiento que da como resultados vectores linealmente separables. O se podría utilizar múltiples perceptrones en múltiples capas. Alternativamente, se puede utilizar otros tipos de redes como redes lineales o redes backpropagation, que puede clasificar vectores de entrada separables no linealmente.

2.4.4. Redes Neuronales Backpropagation

Son RNA supervisadas también denominadas redes de propagación inversa, tienen la misma arquitectura que una red general.

La red Backpropagation se creó generalizando la regla de aprendizaje de Widrow-Hoff para redes de múltiples capas y funciones de transferencia diferenciables no lineales. Los vectores de entrada y los vectores objetivo correspondientes se utilizan para entrenar una red hasta que pueda aproximar una función, asociar vectores de entrada con un vector de salida específico. Las redes con sesgos, una capa sigmoidea

y una capa de salida lineal son capaces de aproximar cualquier función con un número finito de discontinuidades.

El Backpropagation estándar es un algoritmo de descenso por gradiente, al igual que la regla de aprendizaje de Widrow-Hoff, en la que los pesos de la red se mueven a lo largo del negativo del gradiente de la función de desempeño. El termino Backpropagation se refiere a la manera en que se calcula el gradiente para redes multicapa no lineales.

Las redes Backpropagation adecuadamente entrenadas tienden a dar respuestas razonables cuando se les presentan nuevas entradas. Normalmente, una nueva entrada conduce a una salida similar a la salida correcta para los vectores de entrada utilizados en el entrenamiento que son similares a la nueva entrada presentada. La generalización de esta propiedad permite entrenar una red en un conjunto representativo de pares de entrada/salida y obtener buenos resultados sin entrenar la red en todos posibles pares de entrada/salida.

El proceso de entrenamiento de una red Backpropagation consta de cuatro pasos:

a) Cargar los datos de entrenamiento.

b) Crear la red.

c) Entrenar la red.

d) Simular la respuesta de la red a nuevas entradas.

2.4.5. Feedforward

Estas redes frecuentemente tienen una o más capas ocultas de neuronas sigmoide seguidas de una capa de salida de neuronas lineales. Varias capas de las neuronas con funciones de transferencia no lineal permiten que la red aprenda las

relaciones lineales y no lineales entre los vectores de entrada y salida. La salida lineal de la capa permite que la red produzca valores fuera del rango de –1 a +1.

Por otro lado, si desea restringir las salidas de una red (como entre 0 y 1), entonces la capa de salida debe usar una función de transferencia sigmoidea (como logsig).

2.4.6. Redes de Base Radial

Las redes de base radial constan de dos capas: una capa de base radial oculta y una capa lineal de salida. Estas pueden requerir más neuronas que las redes estándar Feed-forward Backpropagation, pero con frecuencia se pueden diseñar en una fracción del tiempo que se tarda en entrenar redes estándar Feed-forward. Estas redes funcionan mejor cuando hay muchos vectores de entrenamiento disponibles.

En una red de base radial, la entrada neta de una neurona es diferente. La entrada neta a la función de transferencia es la distancia vectorial entre su vector de peso *w* y el vector de entrada *p*, multiplicado por el sesgo *b*.

La función de transferencia para una neurona de base radial es:

$$radbas(n) = e^{-n^2}$$

La función de base radial tiene un máximo de 1 cuando su entrada es cero. A medida que la distancia entre *w* y *p* decrece, la salida se incrementa.

2.4.7. Redes Neuronales Probabilísticas

Las redes neuronales probabilísticas se pueden utilizar para problemas de clasificación. Cuando se presenta la entrada, la primera capa calcula las distancias desde el vector de entrada hasta los vectores de entrada de entrenamiento, y produce

un vector cuyos elementos indican cuán cerca está la entrada a la entrada de entrenamiento. La segunda capa suma estas contribuciones para cada clase de entradas para producir como salida neta un vector de probabilidades.

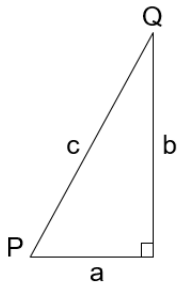
Finalmente, una función de transferencia competitiva en la salida de la segunda capa selecciona el máximo de estas probabilidades, y produce un 1 para esa clase y un 0 para las otras clases. Se supone que existen *Q* pares de vectores de entrada/salida. Cada vector de salida tiene *K* elementos. Uno de estos elementos es 1 y el resto es 0. Así, cada vector de entrada está asociado con una de las *K* clases.

2.4.8. Self-Organizing-Maps

Este tipo de RNA se identifica con la sigla SOM, y su aprendizaje es el no supervisado**.** Las neuronas en la capa de un SOM pueden ser representadas por puntos o posiciones de un plano bidimensional o plano coordenado XY, formando los vértices de un grafo. Los vértices del grafo son las posiciones de las neuronas y las aristas son los enlaces entre estas. Los grafos adoptan diversas formas o geometrías, denominadas Topología. Así tenemos la topología *gridtop* en la que las posiciones de las neuronas en el plano forman una malla o rejilla rectangular, la topología *hextop* en la que la posición de las neuronas en el plano tiene un patrón hexagonal, y finalmente la topología *randtop* en el que las neuronas ocupan una posición aleatoria en el plano.

La distancia entre dos neuronas está determinada por la métrica. En la Figura 19, la distancia Euclidiana, *Dist* de *P* a *Q* es la hipotenusa *c*, la distancia de Manhathan, Mandist de *P* a *Q* es *a*+*b*, la distancia Boxdist es el máximo de los catetos, *b* en este caso, la distancia Linkdist es el enlace más corto y se emplea en las topologías Gridtop y Hextop.

Figura 19. **Distancia entre dos neuronas**



Fuente: Elaboración propia

En cuanto a su arquitectura, es como la de una red cualquiera, pero no emplea sesgos, y sus funciones de transferencia son como las de las RNA competitivas.

Una red Self-Organizing Maps aprende a categorizar los vectores de entrada, debido a que aprende la distribución de los vectores de entrada. Las redes competitivas asignan más neuronas para reconocer partes del espacio de entrada con mayores densidades de entrada, y asignan menos neuronas a partes del espacio de entrada donde ocurren pocos vectores de entrada.

2.5. MÉTODOS HEURÍSTICOS

El término heurística es de origen griego y significa hallar, inventar. Este término se usa como sustantivo y como adjetivo, cuando se usa como sustantivo, se refiere a la ciencia del descubrimiento, y cuando se usa como adjetivo, se refiere a las estrategias y reglas que se siguen en un procedimiento para resolver un problema.

> *Un método heurístico es un procedimiento que trata de descubrir una solución factible muy buena, pero no necesariamente una solución óptima, para el problema específico bajo consideración. No puede darse una garantía acerca de la calidad de la solución que se obtiene, pero un método heurístico bien diseñado puede proporcionar una solución que al menos está cerca de ser óptima (o*

concluir que no existen tales soluciones). El procedimiento también debe ser suficientemente eficiente como para manejar problemas muy grandes. Con frecuencia, el procedimiento es un algoritmo iterativo novedoso, donde cada iteración implica la realización de una búsqueda de una nueva solución que puede ser mejor que la solución que se encontró con anterioridad. Cuando el algoritmo termina después de un tiempo razonable, la solución que proporciona es la mejor que se pudo encontrar en cualquier iteración (Hillier, Lieberman, pág. 563).

La heurística está diseñada para encontrar buenas soluciones aproximadas de problemas combinatorios difíciles que de lo contrario no pueden resolverse mediante los algoritmos de optimización disponibles. Una heurística es una técnica de búsqueda directa que utiliza reglas favorables prácticas para localizar soluciones mejoradas. La ventaja de la heurística es que en general determina (buenas) soluciones con rapidez, utilizando reglas de solución simples. La desventaja es que la calidad de la solución (con respecto a la óptima) suele desconocerse.

Las primeras generaciones de heurística se basan en la regla de búsqueda codiciosa que dicta que se mejore el valor de la función objetivo con cada movimiento de búsqueda. La búsqueda termina en un óptimo local donde ya no son posibles más mejoras.

En la década de 1980, una nueva generación de metaheurística buscó mejorar la calidad de las soluciones heurísticas al permitir la búsqueda de una trampa de escape en óptimos locales (Taha, pág. 351).

En el entrenamiento de las Redes Neuronales y de las Máquinas con Soporte Vectorial intervienen procedimientos heurísticos debido a que estas técnicas de Máquinas de Aprendizaje están basadas en el uso del gradiente que ayuda a determinar un punto crítico de una función no lineal de muchas variables. Es un método heurístico por excelencia porque, el método de gradiente requiere de un punto inicial para ser

ejecutado y luego va conduciendo el proceso hacia un punto que sea el máximo o el mínimo de la función no lineal. Indudablemente estas técnicas han sido mejoradas e implementadas en programas de alta complejidad empleando la matriz Hessiana como ocurre en los artículos de Levenberg (1944).

CAPITULO III: ESTUDIO DE LA BASE DE DATOS

En este capítulo se estudia la Base de datos y se hacen las pruebas empleando técnicas de aprendizaje de máquina. Con las Redes Neuronales Artificiales Backpropagation (RNA) se hace la predicción del comportamiento crediticio de los prestatarios, con las Redes Self-Organizing-Maps (RNA-SOM) se agrupan o categorizan los prestatarios en clústeres, y con las Máquinas con Soporte Vectorial (MSV) se separan los registros de la Base de datos en dos clases. En todos casos se usará Matlab R2018a.

3.1. ESTUDIO ESTADÍSTICO DE LA BASE DE DATOS

3.1.1. Datos y variables

Para el presente estudio se tiene una Base de datos (BD) de 15,569 registros de prestatarios; cada registro consta de 26 variables. La variable V27 es la variable dependiente e indica la aceptación (0) o el rechazo (1) del crédito. En el entorno de Matlab, BD es una matriz de 15,569 filas (registros) por 27 columnas (variables). Las variables conjuntamente con su descripción se muestran en la Tabla 5 y en el Anexo 1 se muestra el Diccionario de datos.

Tabla 5. Variables de la Base de datos

Variable	Descripción
V1	Tipo de moneda.
V2	Monto del crédito otorgado.
V3	Saldo capital de la deuda.
V4	Tipo de Crédito según Reporte Crediticio de Deudores.
V5	Clasificación del deudor.
V6	Clasificación del deudor sin considerar alineamiento con el sistema.
V7	Días de atraso al cierre del mes.
V8	Días de atraso de la última cuota pagada.
V9	Promedio de días de atraso en el pago de cuotas en los últimos 6 meses.
V10	Provisión constituida.
V11	Saldo de capital vigente de la operación.

V12	Saldo de capital vencido de la operación.
V13	Saldo de capital en cobranza judicial de la operación.
V14	Rendimientos devengados de la operación.
V15	Intereses en suspenso acumulados de la operación.
V16	Fecha de desembolso.
V17	Esquema de amortización.
V18	Número de días de gracia para pago de capital según cronograma.
V19	Fecha de vencimiento general de la operación.
V20	Fecha de vencimiento puntual de la operación.
V21	Periodicidad de cuotas.
V22	Número de cuotas programadas.
V23	Número de cuotas pagadas.
V24	Indicador de Rescate Financiero Agropecuario.
V25	Código de agencia.
V26	Tasa efectiva anual.
V27	Aceptación o rechazo del crédito (0: Se acepta, 1: Se rechaza).

Fuente: Elaboración propia

3.1.2. Medidas estadísticas y coeficiente de correlación

En primer lugar, se calculan las medidas estadísticas mínimo, máximo, desviación estándar, media, moda y el coeficiente de correlación de la variable Aceptación o rechazo del crédito (V27) con las primeras 26 variables. En la Tabla 6 se muestra esta información.

Tabla 6. Medidas estadísticas y coeficiente de correlación de la Base de datos

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.03	1	1	-0.01
V2	300	297770	10957.9	5879.45	1000	0.01
V3	16.24	297770	10475.32	5078.65	1000	0.01
V4	8	13	1.1	11.07	12	-0.03
V5	0	4	0.59	0.15	0	0.43
V6	0	4	0.55	0.13	0	0.45

V7	-419	268	37.64	-14.98	-15	0.26
V8	**0**	**162**	**8.32**	**2.88**	**0**	**0.78**
V9	-125	88	4.6	0.02	0	0.21
V10	0.83	48000	756.33	136.1	20	0.13
V11	0	297770	10480.37	5034.86	1000	0
V12	0	26527	453.83	38.05	0	0.22
V13	0	29822	350.02	5.75	0	0.04
V14	0	19652.81	498.09	164.71	0	0.05
V15	0	5903	94.62	9.03	0	0.19
V16	40546	40847	84.95	40711.96	40751	-0.14
V17	1	5	0.35	2.94	3	0.03
V18	0	360	21.54	6.35	0	0.03
V19	40714	48131	438.13	41198.31	41001	-0.01
V20	40579	41209	37.64	40862.16	40849	-0.26
V21	1	365	35.29	35.84	30	-0.02
V22	1	240	14.3	15.65	12	0.01
V23	0	37	2.77	3.47	0	0.03
V24	1	1	0	1	1	NaN
V25	1	44	13.38	18.98	1	-0.01
V26	12.01	293.79	18.25	51.09	58.27	0.04
V27	0	1	0.14	0.02	0	1

Fuente: Elaboración propia

Se observa que la variable Días de atraso de la última cuota pagada (V8) tiene el coeficiente de correlación más alto (0.78) con la variable V27. Esta variable es la respuesta real del sistema que está usando la entidad de micro-finanzas, porque está dando una variabilidad de respuesta. Por tanto, se puede afirmar que la variable V27 (con valores 0 y 1) es artificial, porque depende de V8 y anula un estudio inteligente de la Base de datos, debido a que cuando esta variable tiene valores entre 0 y 30 días, la variable V27 tiene valor 0, y cuando V8 tiene valores de 31 días a más, V27 tiene valor 1. Es decir, la variable V27 es totalmente compresora de datos.

Entonces, luego de este análisis, una conclusión es eliminar la variable V27 y reemplazarla por V8, la Tabla 7 muestra el coeficiente de correlación que tienen las 25

variables con V8, en ella se aprecia que las variables que tenían un coeficiente de correlación significativo con V27, lo siguen teniendo con V8. Este hallazgo se tendrá en cuenta para las pruebas con las Redes Neuronales Artificiales (RNA) y con las Máquinas con Soporte Vectorial (MSV).

Tabla 7. Coeficientes de correlación de las 25 variables con V8

Var	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**	**V7**	**V9**	**V10**
CC	-0.01	0	0	-0.04	0.51	0.53	0.36	0.33	0.14
Var	**V11**	**V12**	**V13**	**V14**	**V15**	**V16**	**V17**	**V18**	**V19**
CC	-0.01	0.24	0.03	0.05	0.22	-0.24	0.06	0.03	-0.03
Var	**V20**	**V21**	**V22**	**V23**	**V24**	**V25**	**V26**	**V8**	
CC	-0.36	-0.06	0.01	0.10	NaN	-0.03	0.07	1	

Fuente: Elaboración propia

3.2. ESTUDIO CON REDES NEURONALES ARTIFICIALES (RNA)

Se desarrollará una RNA Backpropagation supervisada, teniéndose como entrada de la red, la Base de datos (BD) referida en la sección 3.1.1 para pronosticar el comportamiento crediticio del prestatario. Se harán pruebas con la BD completa, luego con la BD en la que la variable V8 reemplaza a la variable V27, y finalmente con la BD en la que se elimina la variable V8. En cada caso se probarán diferentes arquitecturas de la red neuronal. Las gráficas han sido obtenidas con Matlab versión R2018a.

3.2.1. Pruebas con la base de datos completa

Se trabaja con la Base de datos referida en la sección 3.1.1 que contiene 15,569 registros cada uno con 27 variables. En el entorno de Matlab, la Base de datos es una matriz de 15,569 filas (registros) por 27 columnas (variables).

Procedimiento para la creación de la red:

(i) Separación de los datos y del valor esperado.

(ii) Separación de datos para el pronóstico de la red.

(iii) Normalización y reducción de las variables altamente correlacionadas mediante la Técnica de las Componentes Principales.

(iv) Separación de las columnas para el entrenamiento, validación y test de la red.

(v) Entrenamiento.

(vi) Validación.

(vii) Test.

a) Prueba con una RNA de cuatro capas con 14, 10, 8 y 1 neuronas respectivamente

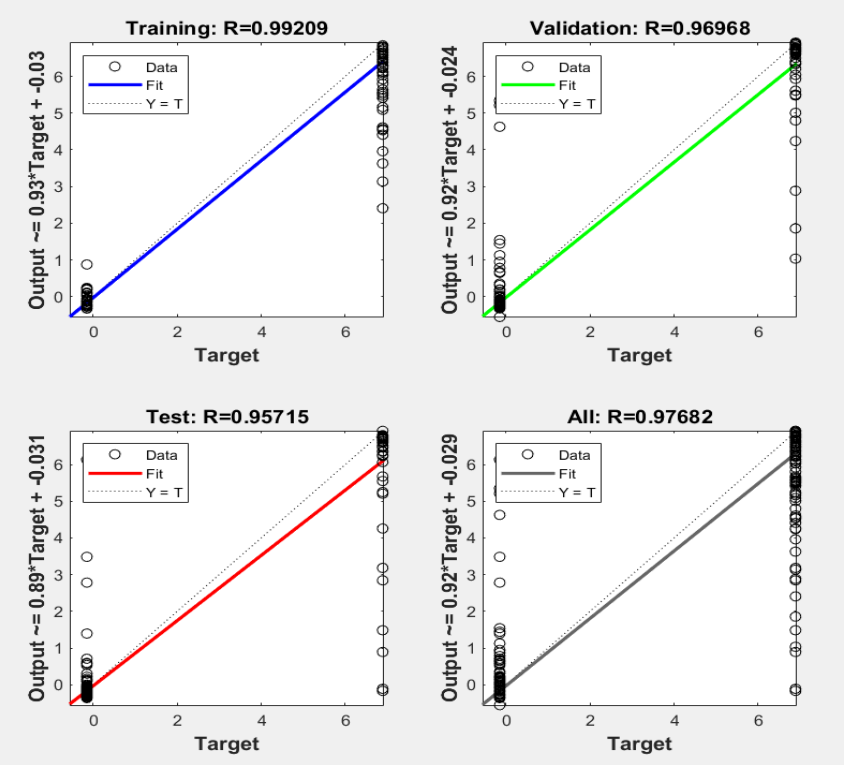
Procedimiento:

(i) Creación de la red con la arquitectura indicada.

(ii) Entrenamiento de la red.

Se ejecutan 20 iteraciones y se obtienen los resultados que se muestran en la Figura 20.

Figura 20. **Valores de Precisión obtenidos en la prueba**



Fuente: Obtenido con Matlab R2018a

b) Prueba con una RNA de cuatro capas con 20, 14, 8 y 1 neuronas respectivamente

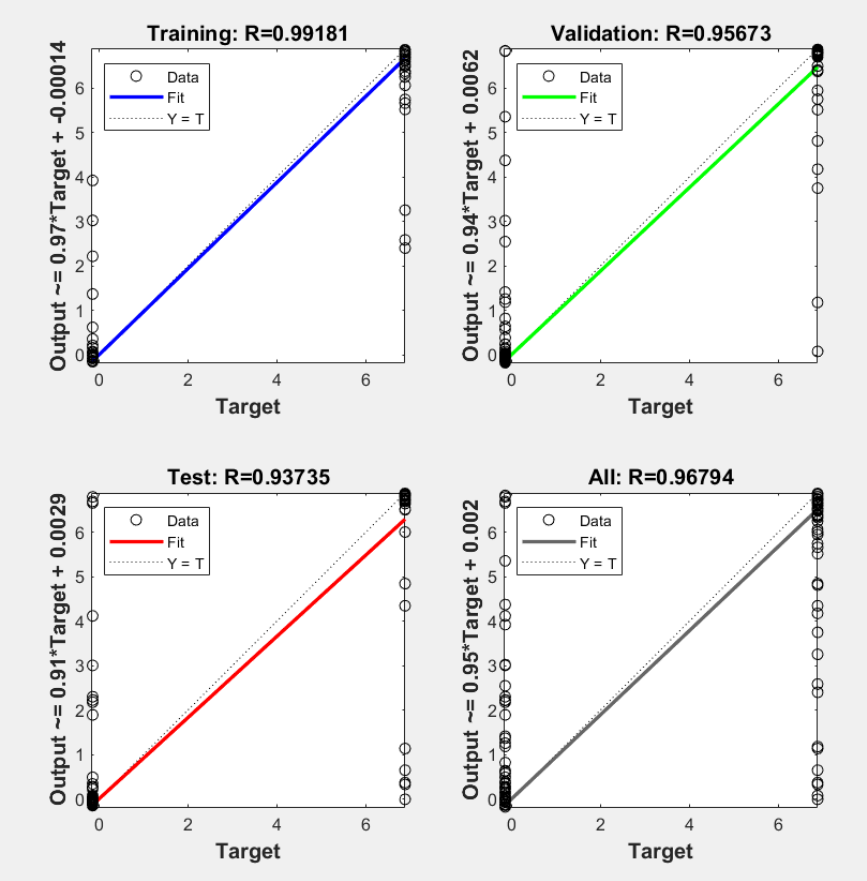
Procedimiento:

(i) Creación de la red con la arquitectura indicada.

(ii) Entrenamiento de la red.

Se ejecutan 22 iteraciones y se obtienen los resultados que se muestran en la Figura 21.

Figura 21. **Valores de Precisión obtenidos en la prueba**



Fuente: Obtenido con Matlab R2018a

c) Prueba con una RNA de tres capas con 20, 10 y 1 neuronas respectivamente

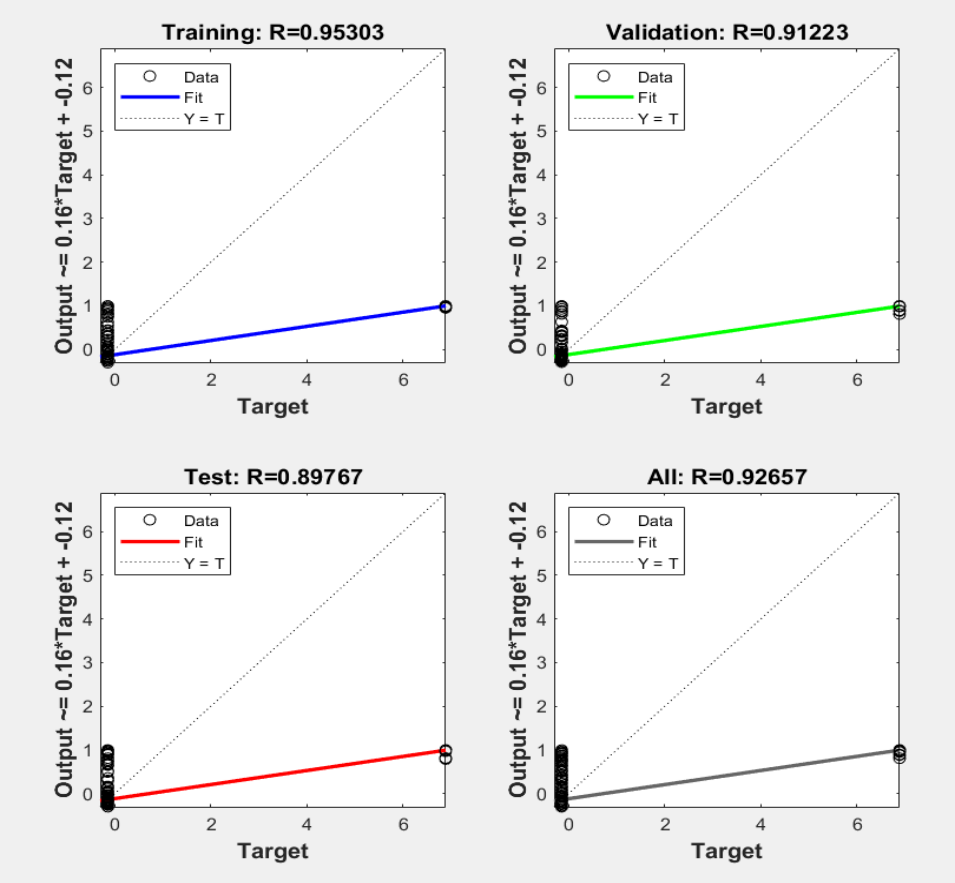
Procedimiento:

(i) Creación de la red con la arquitectura indicada.

(ii) Entrenamiento de la red.

Se ejecutan 48 iteraciones y se obtienen los resultados que se muestran en la Figura 22.

Figura 22. **Valores de Precisión obtenidos en la prueba**



Fuente: Obtenido con Matlab R2018a

d) Prueba con una RNA de cinco capas con 20, 18, 14, 10 y 1 neuronas respectivamente

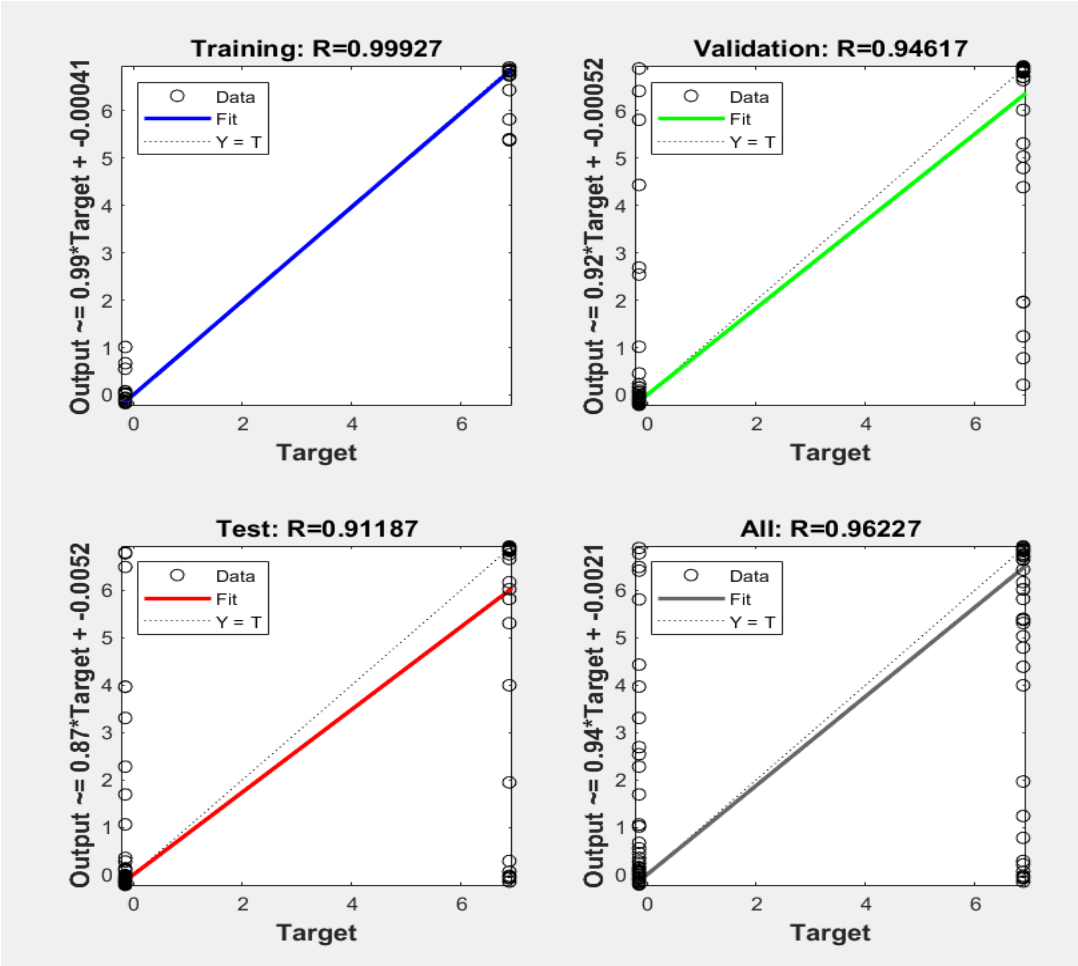
Procedimiento:

(i) Creación de la red con la arquitectura indicada.

(ii) Entrenamiento de la red.

Se ejecutan 25 iteraciones y se obtienen los resultados que se muestran en la Figura 23.

Figura 23. **Valores de Precisión obtenidos en la prueba**



Fuente: Obtenido con Matlab R2018a

3.2.2. Pruebas con la base de datos con 26 variables (V8 reemplaza a V27)

Debido a que la variable Días de atraso de la última cuota pagada (V8) tiene la correlación más alta (0.78) con la variable Aceptación o rechazo del crédito (V27), se puede afirmar que hay una dependencia entre estas dos variables, es decir, la variable V27 depende de la variable V8, por esta razón, en estas pruebas se elimina V27 y se reemplaza por V8. En el entorno de Matlab, la Base de datos (BD) es una matriz de 15,569 filas (registros) por 26 columnas (variables).

Procedimiento para la creación de la red:

(i) Separación de los datos y del valor esperado.

(ii) Separación de datos para el pronóstico de la red.

(iii) Normalización y reducción de las variables altamente correlacionadas mediante la Técnica de las Componentes Principales.

(iv) Separación de las columnas para el entrenamiento, validación y test de la red.

(v) Entrenamiento.

(vi) Validación.

(vii) Test.

a) Prueba con una RNA de cuatro capas con 14, 10, 8 y 1 neuronas respectivamente

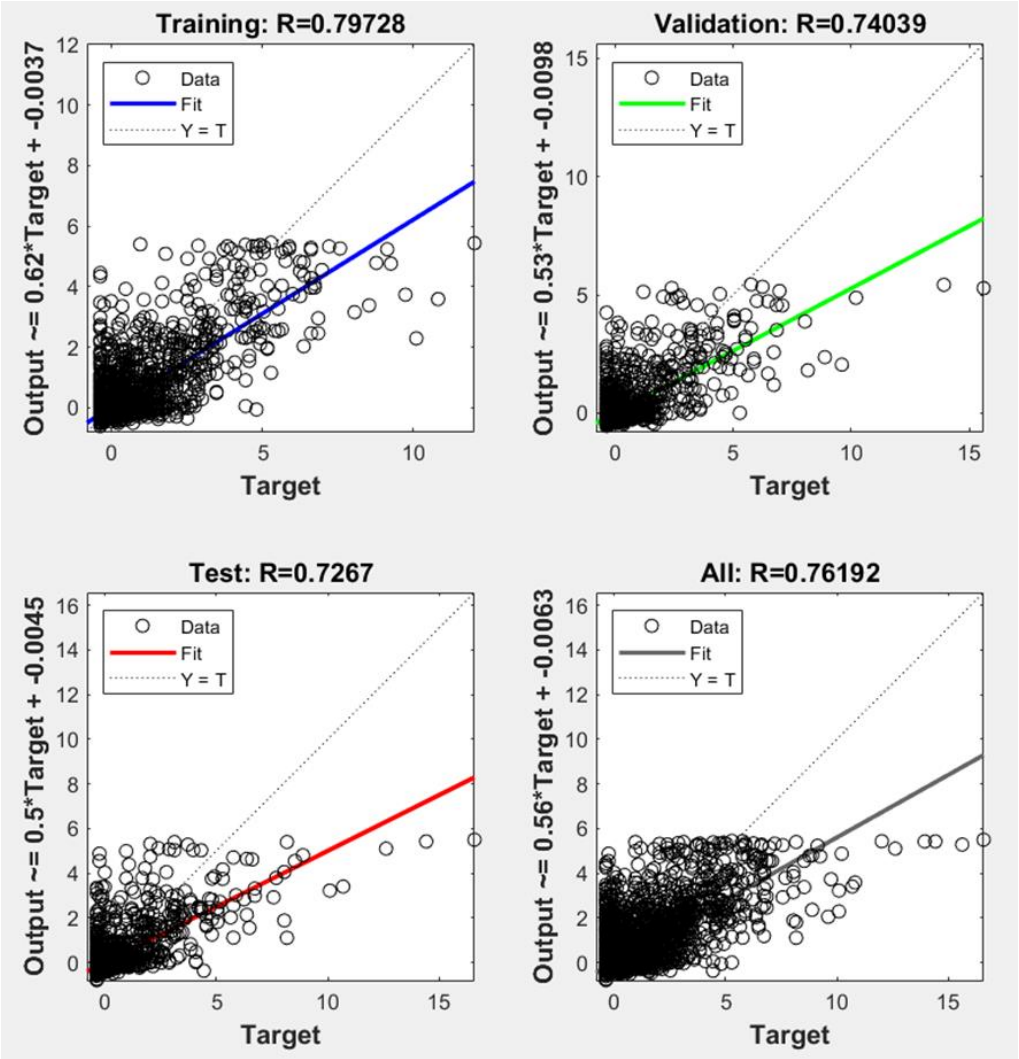
Procedimiento:

(i) Creación de la red con la arquitectura indicada.

(ii) Entrenamiento de la red.

Se ejecutan 18 iteraciones con los resultados que se muestran en la Figura 24.

Figura 24. **Valores de Precisión obtenidos en la prueba**



Fuente: Obtenido con Matlab R2018a

b) Prueba con una RNA de cuatro capas con 20, 14, 8 y 1 neuronas respectivamente

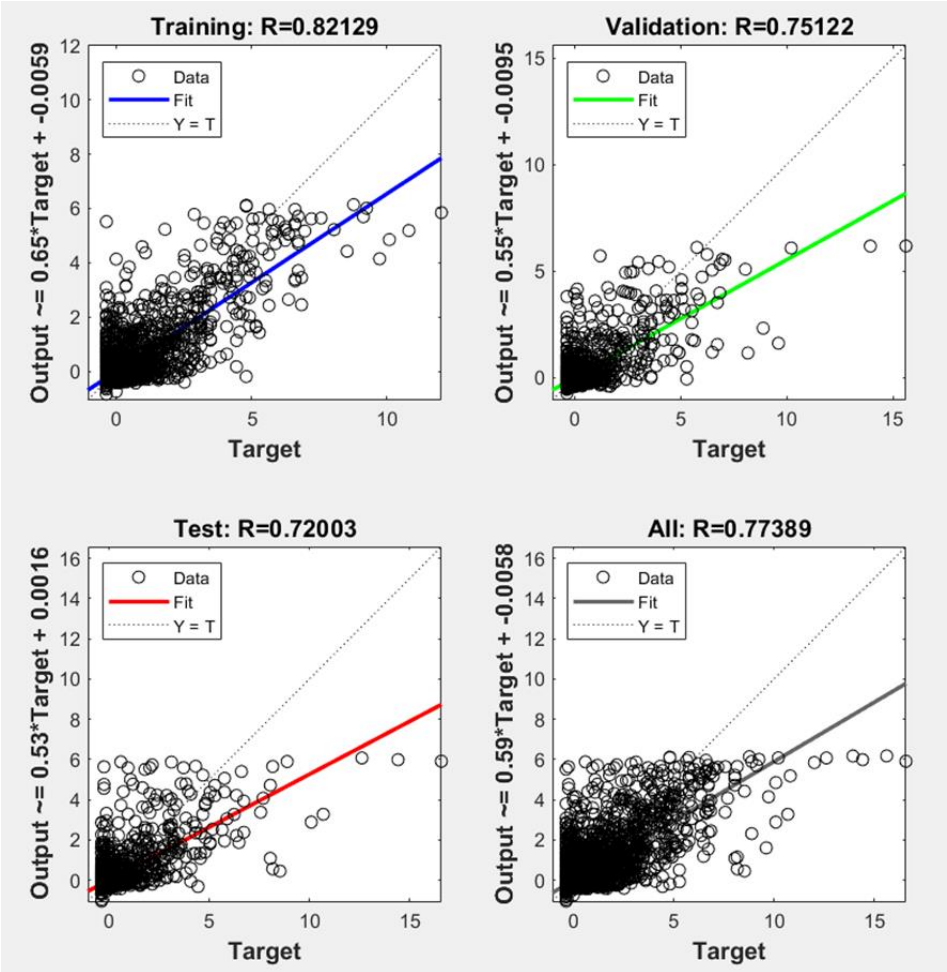
Procedimiento:

(i) Creación de la red con la arquitectura indicada.

(ii) Entrenamiento de la red.

Se ejecutan 18 iteraciones y se obtienen los resultados que se muestran en la Figura 25. con esta arquitectura se obtiene la mejor correlación global: 0.77389

Figura 25. **Valores de Precisión obtenidos en la prueba**



Fuente: Obtenido con Matlab R2018a

c) Prueba con una RNA de tres capas con 20, 10 y 1 neuronas respectivamente

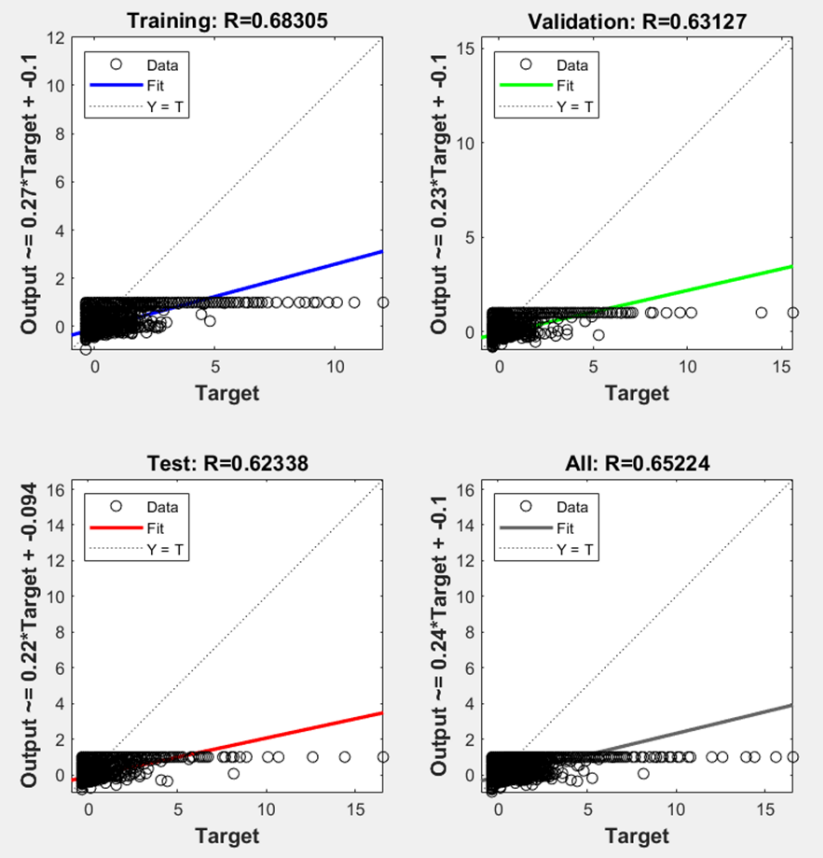
Procedimiento:

(i) Creación de la red con la arquitectura indicada.

(ii) Entrenamiento de la red.

Se ejecutan 28 iteraciones y se obtiene los resultados que se muestran en la Figura 26.

Figura 26. **Valores de Precisión obtenidos en la prueba**



Fuente: Obtenido con Matlab R2018a

d) Prueba con una RNA de cinco capas con 20, 18, 14, 10 y 1 neuronas respectivamente

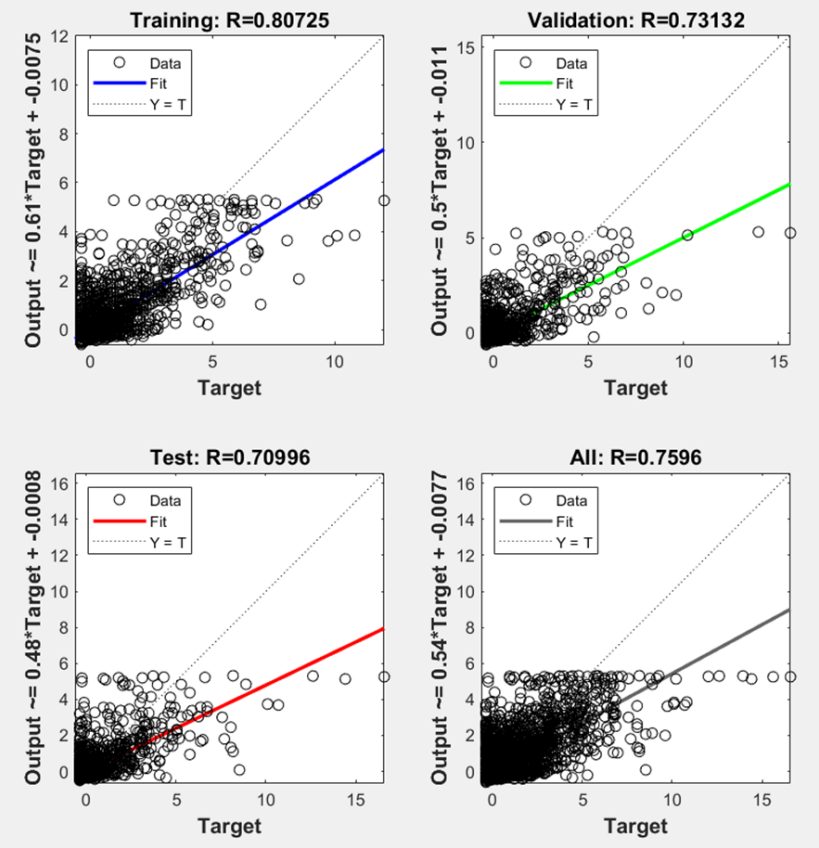
Procedimiento:

(i) Creación de la red con la arquitectura indicada.

(ii) Entrenamiento de la red.

Se ejecutan 15 iteraciones y se obtienen los resultados que se muestran en la Figura 27.

Figura 27. **Valores de Precisión obtenidos en la prueba**



Fuente: Obtenido con Matlab R2018a

3.2.3. Pruebas con la base de datos con 26 variables (se elimina V8)

En esta prueba, se elimina de la base de datos la variable Días de atraso de la última cuota pagada (V8) que tiene la correlación más alta (0.78) con la variable Aceptación o rechazo del crédito (V27). En el entorno de Matlab, BD es una matriz de 15,569 filas (registros) por 26 columnas (variables).

Procedimiento para la creación de la red:

(i) Separación de los datos y del valor esperado.

(ii) Separación de datos para el pronóstico de la red.

(iii) Normalización y reducción de las variables altamente correlacionadas mediante la Técnica de las Componentes Principales.

(iv) Separación de las columnas para el entrenamiento, validación y test de la red.

(v) Entrenamiento.

(vi) Validación.

(vii) Test.

a) Prueba con una RNA de cuatro capas con 14, 10, 8 y 1 neuronas respectivamente

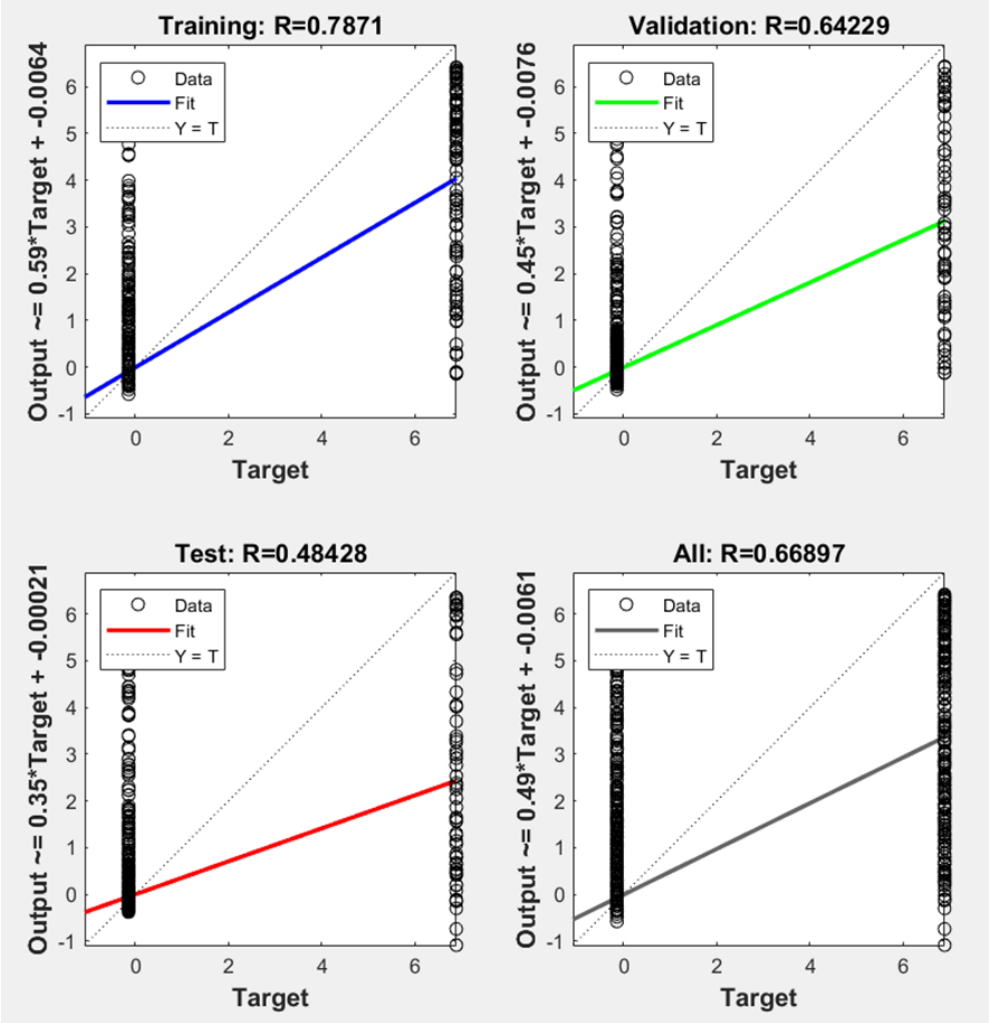
Procedimiento:

(i) Creación de la red con la arquitectura indicada.

(ii) Entrenamiento de la red.

Se ejecutan 15 iteraciones y se obtienen los resultados que se muestran en la Figura 28.

Figura 28. **Valores de Precisión obtenidos en la prueba**



Fuente: Obtenido con Matlab R2018a

b) Prueba con una RNA de cuatro capas con 20, 14, 8 y 1 neuronas respectivamente

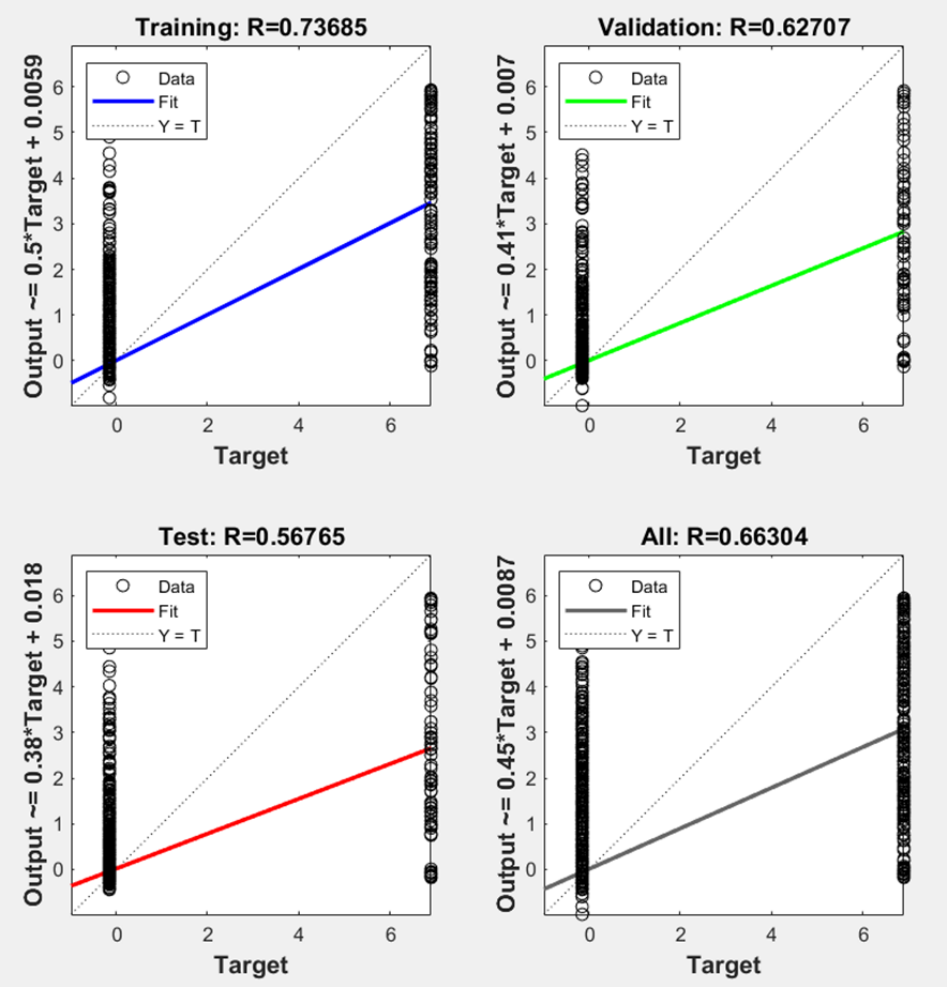
Procedimiento:

(i) Creación de la red con la arquitectura indicada.

(ii) Entrenamiento de la red.

Se ejecutan 17 iteraciones y se obtienen los resultados que se muestran en la Figura 29.

Figura 29. **Valores de Precisión obtenidos en la prueba**



Fuente: Obtenido con Matlab R2018a

c) Prueba con una RNA de tres capas con 20, 10 y 1 neuronas respectivamente

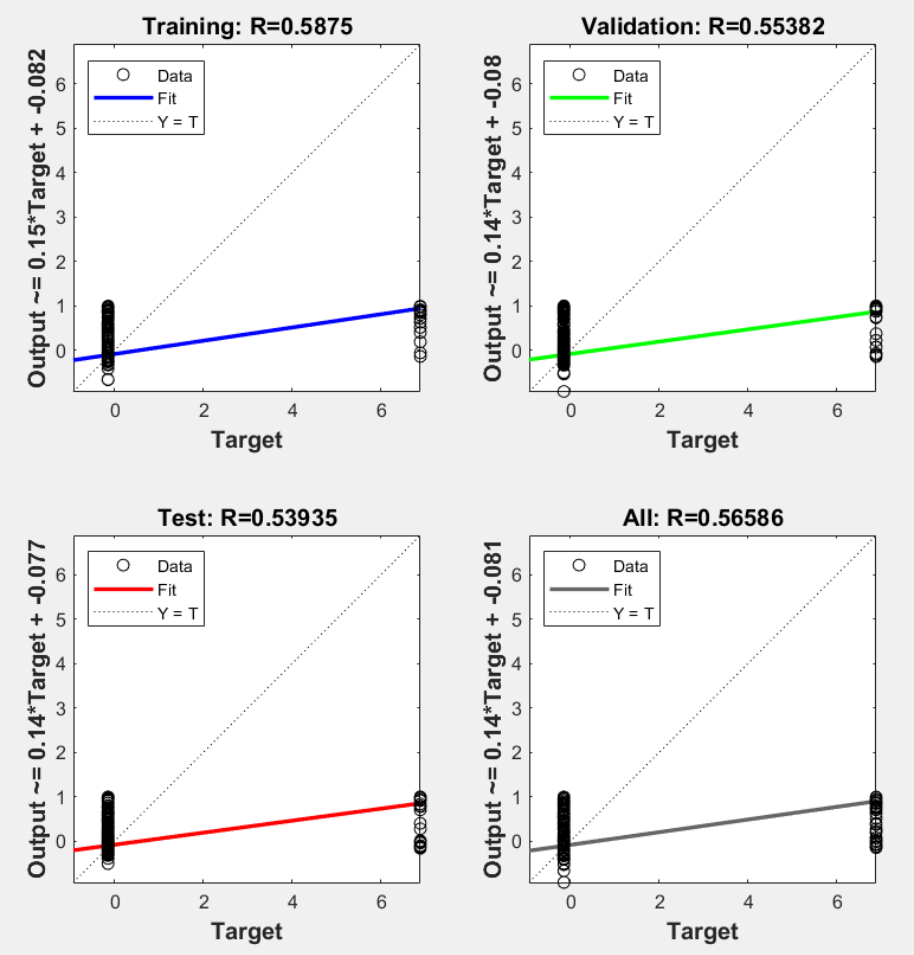
Procedimiento:

(i) Creación de la red con la arquitectura indicada.

(ii) Entrenamiento de la red.

Se ejecutan 24 iteraciones y se obtienen los resultados que se muestran en la Figura 30.

Figura 30. **Valores de Precisión obtenidos en la prueba**



Fuente: Obtenido con Matlab R2018a

d) Prueba con una RNA de cinco capas con 20, 18, 14, 10 y 1 neuronas respectivamente

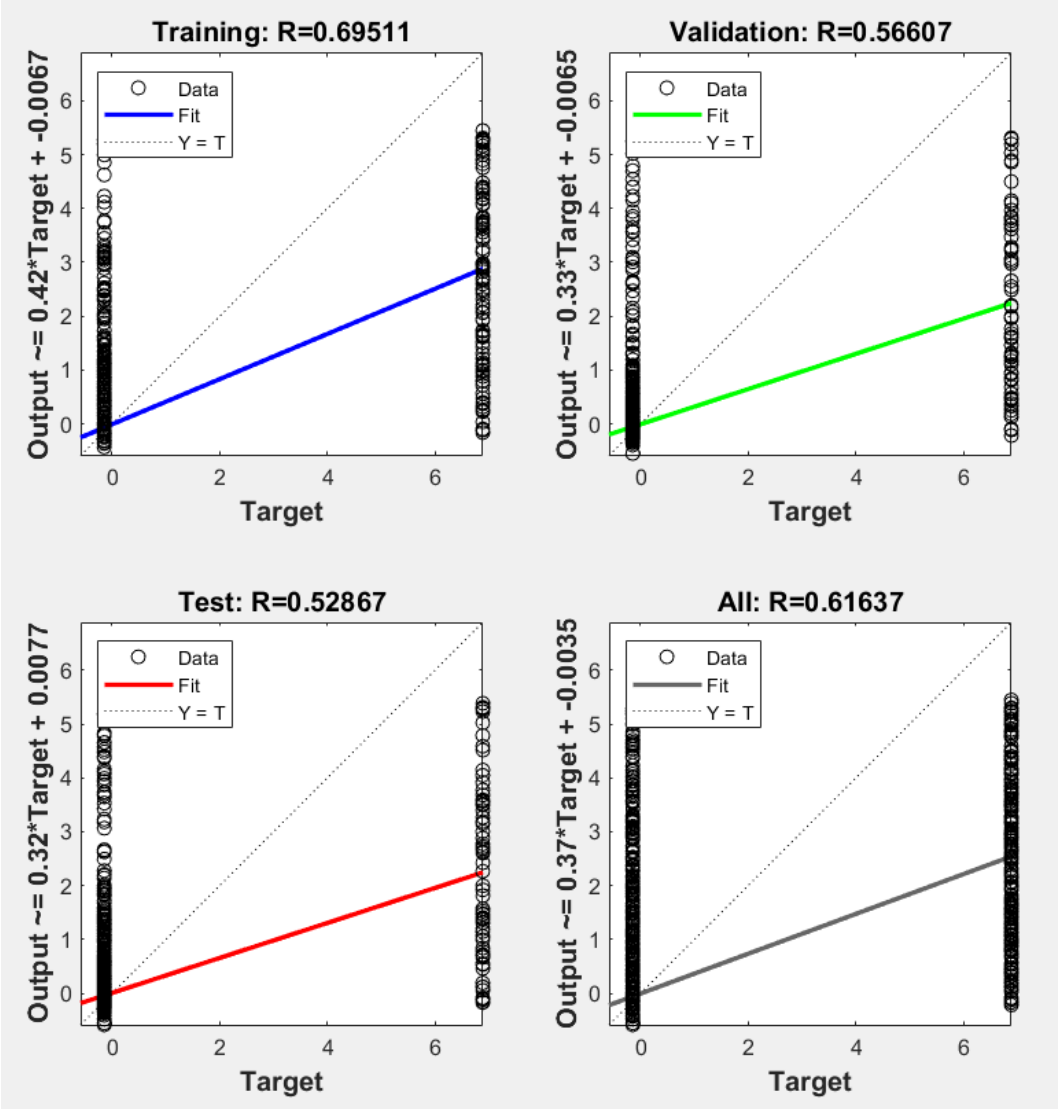
Procedimiento:

(i) Creación de la red con la arquitectura indicada.

(ii) Entrenamiento de la red.

Se ejecutan 13 iteraciones y se obtienen los resultados que se muestran en la Figura 31.

Figura 31. **Valores de Precisión obtenidos en la prueba**



Fuente: Obtenido con Matlab R2018a

3.3. ESTUDIO CON REDES NEURONALES SELF-ORGANIZING-MAPS (SOM)

En esta sección se hacen las pruebas con las Redes Neuronales SOM con dos y tres neuronas, con diferentes topologías y métricas.

3.3.1. Red Neuronal SOM: 2 Neuronas, Gridtop, Dist

Esta red tiene 2 neuronas, usa la topología Gridtop y la métrica Dist.

a) Ingreso de datos y creación de la Red Neuronal SOM

BD es una matriz de 15569 filas por 27 columnas. Las filas son los datos concernientes a cada cliente y las columnas son las variables, las 26 primeras columnas son los datos de información de los clientes y la última columna registra la aceptación (0) o el rechazo del crédito (1).

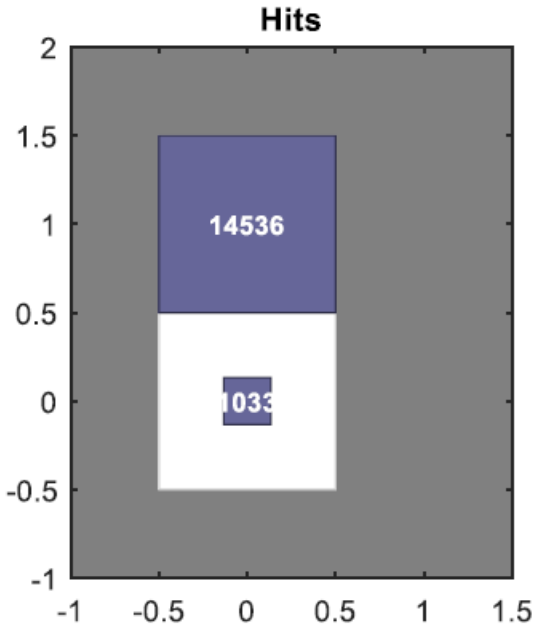
Red neuronal SOM de 2 neuronas

El propósito es generar dos clústeres, solamente con los datos de los clientes, esto es con la matriz P. La matriz T no interviene en el entrenamiento de la red SOM. Para esto se genera una red SOM con 2 neuronas [1 2] con la función selforgmap().

El entrenamiento

Para el entrenamiento de la red se le entrega únicamente la matriz P. Los clústeres creados se muestran en la Figura 32.

Figura 32. **Clústeres creados por la Red SOM**



Fuente: Obtenido con Matlab R2018a

b) Los clústeres y sus miembros

La red SOM ha creado 2 clústeres. Se verá exactamente qué miembros de los datos de ingreso (columnas de P) son los que conforman los clústeres, en otros términos, equivale a conocer qué tipo de clientes están en el clúster 1 y quienes en el clúster 2. Ello se descubrirá estudiando cada una de estas clases.

En la Figura 32 el clúster 1 tiene 1033 registros y el clúster 2 tiene 14536 registros. Este caso, en ambos clústeres hay elementos comunes: aceptados y rechazados.

Los miembros concretos de estos clústeres se consiguen con el código correspondiente.

c) Variables influyentes en la formación de los clústeres

Para identificar las variables que más han influido en la formación de los clústeres se determina el coeficiente de correlación de las 27 variables con el clúster.

d) Separación de los clústeres

Luego de ejecutar el código correspondiente, se obtienen los dos clústeres. El clúster C1 tiene 1,033 prestatarios y el clúster C2 tiene 14,536 prestatarios. La agrupación que ha hecho la Red SOM es para clasificar de acuerdo a los datos de los clientes en 2 grupos. Se esperaba encontrar 2 clústeres, una de rechazados y otra de aceptados, pero no se ha encontrado esa clasificación, en cada clúster existen aceptados y rechazados.

3.3.2. Red Neuronal SOM: 2 Neuronas, Hextop, Linkdist

Esta red tiene 2 neuronas, usa la topología Hextop y la métrica Linkdist

a) Ingreso de datos y creación de la Red Neuronal SOM

BD es una matriz de 15569 filas por 27 columnas. Las filas son los datos concernientes a cada cliente y las columnas son las variables. Como se sabe las 26 primeras columnas son los datos de información de los clientes y la última columna registra la aceptación (0) o el rechazo del crédito (1).

Red neuronal SOM de 2 neuronas

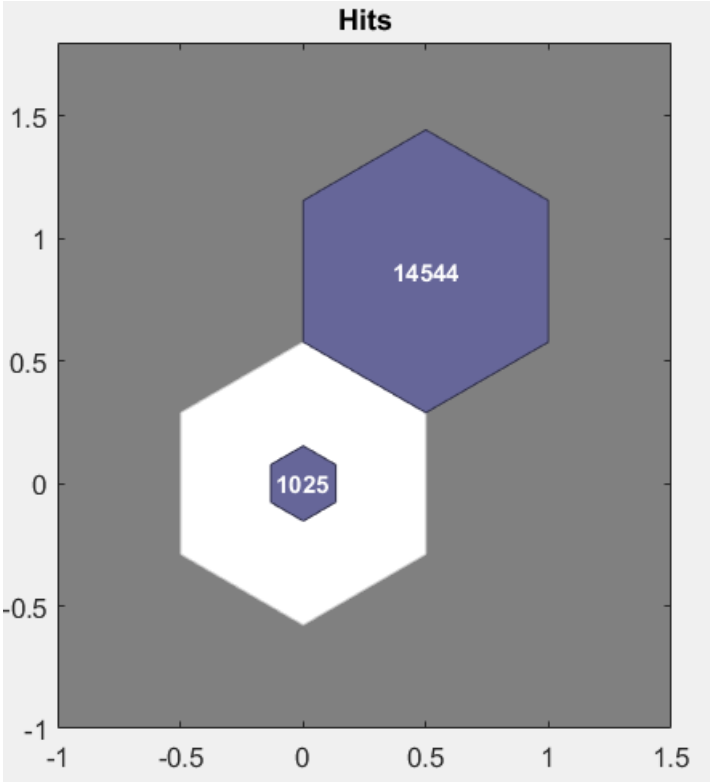
El propósito es generar dos clústeres, solamente con los datos de los clientes, esto es con la matriz P. La matriz T no interviene en el entrenamiento de la red SOM. Para esto se genera una red SOM con 2 neuronas [1 2] con la topología Hextop y la distancia Linkdist (todos los parámetros se ejecutan por defecto en MatLab). Se usa la función *selforgmap().*

El entrenamiento

Luego entrenamos la red, entregándole únicamente la matriz P.

Se obtiene la Figura 33 que muestra los dos clústeres.

Figura 33. **Clústeres creados por la red SOM**



Fuente: Obtenido con Matlab R2018a

b) Los clústeres y sus miembros

La red SOM ha creado 2 clústeres. Se verá exactamente qué miembros de los datos de ingreso (columnas de P) son los que conforman los clústeres, en otros términos, equivale a conocer qué tipo de clientes están en el clúster 1 y quienes en el clúster 2. Ello se descubrirá estudiando cada una de estas clases.

En la gráfica el clúster 1 tiene 1,025 registros y el clúster 2 tiene 14,544 registros. Este caso, en ambos clústeres hay elementos comunes: aceptados y rechazados. Los miembros de cada clase se obtienen con el código correspondiente.

c) Variables influyentes en la formación de los clústeres

Para identificar las variables que más han influido en la formación de los clústeres se determina el coeficiente de correlación de las 27 variables con el clúster.

d) Separación de los clústeres

Luego de ejecutar el código correspondiente, se obtienen los dos clústeres. El clúster C1 tiene 1,025 prestatarios y el clúster C2 tiene 14,544 prestatarios. La agrupación que ha hecho la Red SOM es para clasificar de acuerdo a los datos de los clientes en 2 grupos. En este caso, en cada clúster también existen aceptados y rechazados.

3.3.3. Red Neuronal SOM: 3 Neuronas, Gridtop, Dist

Esta red tiene 3 neuronas, usa la topología Gridtop y la métrica Dist.

a) Ingreso de datos y creación de la Red Neuronal SOM

BD es una matriz de 15,569 filas por 27 columnas. Las filas son los datos concernientes a cada cliente y las columnas son las variables. Como se sabe las 26 primeras columnas son los datos de información de los clientes y la última columna registra la aceptación (0) o el rechazo del crédito (1).

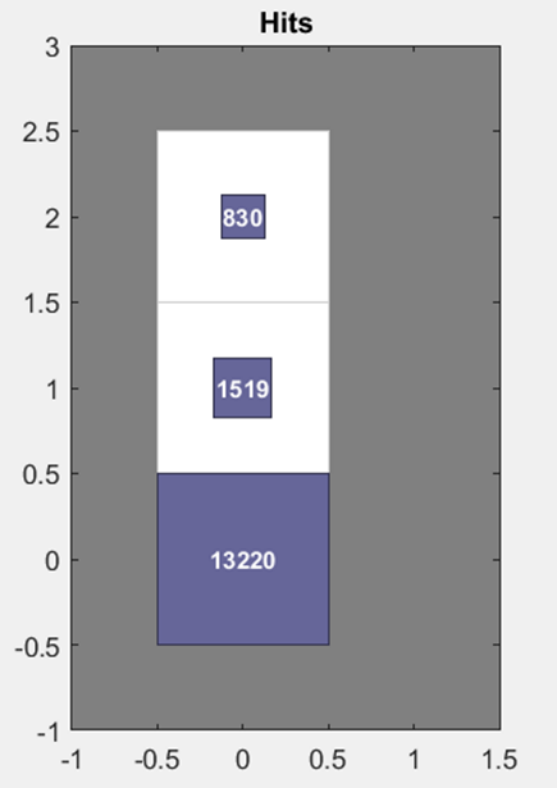
Red neuronal SOM de 3 neuronas

El propósito es generar tres clústeres, solamente con los datos de los clientes, esto es con la matriz P. La matriz T no interviene en el entrenamiento de la red SOM. Para esto se genera una red SOM con 3 neuronas [1 3] con la función *selforgmap().*

El entrenamiento

Luego entrenamos la red, entregándole únicamente la matriz P. Se obtiene los clústeres de la Figura 34.

Figura 34. **Clústeres creados por la red SOM**



Fuente: Obtenido con Matlab R2018a

b) Los clústeres y sus miembros

La red SOM ha creado 3 clústeres. En la Figura 34 el clúster 1 tiene 13,220 registros, el clúster 2 tiene 1,519 registros y el clúster 3 tiene 830 registros. En este caso, en ambos clústeres hay elementos comunes: aceptados y rechazados. Se verá

exactamente qué miembros de los datos de ingreso (columnas de P) son los que conforman los clústeres, en otros términos, equivale a conocer qué tipo de clientes están en cada clúster. Ello se descubrirá estudiando cada una de estas clases.

Se ejecuta el código correspondiente para obtener los miembros de cada clúster.

c) Variables influyentes en la formación de los clústeres

Para identificar las variables que más han influido en la formación de los clústeres se determina el coeficiente de correlación de las 27 variables con el clúster.

d) Separación de los clústeres

Luego de ejecutar el código correspondiente, se obtienen los tres clústeres. El clúster C1 tiene 13,220 prestatarios, el clúster C2 tiene 1,519 prestatarios y el clúster C3 tiene 830 prestatarios. En cada clúster existen aceptados y rechazados, se estudiará la composición de cada clúster.

3.3.4. Red Neuronal SOM: 3 Neuronas, Hextop, Linkdist

Esta red tiene 3 neuronas, usa la topología Hextop y la métrica LinkDist.

a) Ingreso de datos y creación de la Red Neuronal SOM

BD es una matriz de 15569 filas por 27 columnas. Las filas son los datos concernientes a cada cliente y las columnas son las variables. Como se sabe las 26 primeras columnas son los datos de información de los clientes y la última columna registra la aceptación (0) o el rechazo del crédito (1).

Red neuronal SOM de 3 neuronas

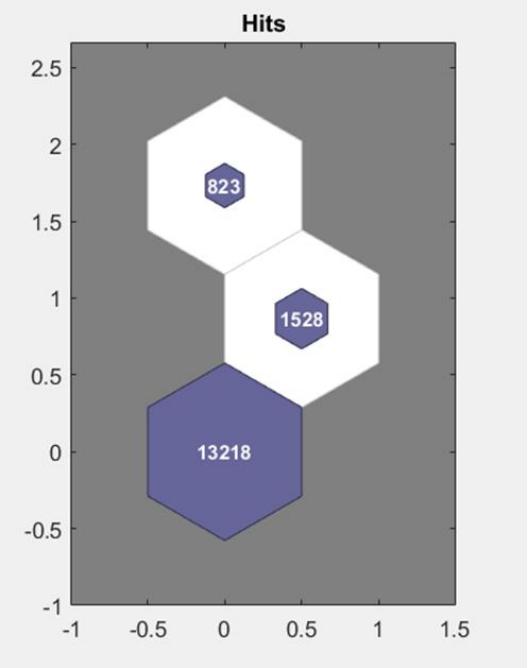
El propósito es generar tres clústeres, solamente con los datos de los clientes, esto es con la matriz P. La matriz T no interviene en el entrenamiento de la red SOM. Se

genera una red SOM con 3 neuronas [1 3] con la topología Hextop y la distancia Linkdist. Se usa la función *selforgmap().*

El entrenamiento

Se entrena la red entregándole únicamente la matriz P. Se obtiene los clústeres mostrados en la Figura 35.

Figura 35. **Clústeres creados por la red SOM**



Fuente: Obtenido con Matlab R2018a

b) Los clústeres y sus miembros

La red SOM ha creado 3 clústeres. Se determinará los miembros de los datos de ingreso (columnas de P) que conforman los clústeres, es decir, se debe conocer qué tipo

de clientes están en el clúster 1, quienes están en el clúster 2, y quienes en el clúster 3. Ello se descubrirá estudiando cada una de estas clases.

En la Figura 35 el clúster 1 tiene 13,218 registros, el clúster 2 tiene 1,528 registros, y el clúster 3 tiene 823 registros.

Se ejecuta el código correspondiente para obtener los miembros de los clústeres.

c) Variables influyentes en la formación de los clústeres

Para identificar las variables que más han influido en la formación de los clústeres se determina el coeficiente de correlación de las 27 variables con el clúster.

d) Separación de los clústeres

Con el código correspondiente se separan los 3 clústeres. El clúster C1 tiene 13,220 prestatarios, el clúster C2 tiene 1,519 prestatarios y el clúster C3 tiene 830 prestatarios. En cada clúster existen aceptados y rechazados, se estudiará que registros conforman cada clúster.

3.4. ESTUDIO CON MÁQUINAS CON SOPORTE VECTORIAL (MSV)

La base de datos BD tiene 15569 registros, 26 variables de información de los prestatarios y la variable V27 indica la aceptación (+1) o rechazo (-1) del crédito. Entonces en el modelo matemático: $A = \{a_1, a_2, ..., a_m\} \subset \Re^n$, m=15569, n=26 y las categorías $y_1, y_2, ..., y_m$, donde $y_i = \pm 1$ son los valores de V27. El conjunto de pares $(a_1, y_1), (a_2, y_2), ..., (a_m, y_m)$ son los datos de entrenamiento.

Al usar métodos lineales de clasificación el problema es separar el conjunto A en dos clases disjuntas:

A(+) = {a_i, con y_i=1} y A(–) = {a_i, con y_i = –1}

Mediante un hiperplano H de separación óptima. Esta afirmación significa que H está entre dos hiperplanos paralelos H(+) y H(–) que tienen la máxima separación, determinados por vectores de A(+) y A(–) respectivamente y denominados vectores soporte. En consecuencia la región limitada por H(+) y H(–) no contiene puntos de A, pero contiene a H que es equidistante de los hiperplanos frontera. Desde el punto de vista matemático los hiperplanos H, H(+) y H(-) son paralelos o coinciden o se interceptan formando un espacio afín de dimensión menor, no queda otra posibilidad.

Un hiperplano en $\Re^n$ de coeficientes $c = (c_1, c_2, ..., c_n)$ y termino independiente b es el conjunto de puntos $x = (x_1, x_2, ..., x_n)$ que satisfacen la ecuación:

L: $c_1 x_1 + c_2 x_2 + ... + c_n x_n + b = 0$; brevemente L: $< x, c > + b = 0$

La distancia (con signo) de un punto u al hiperplano L está dada por

$$d = \frac{1}{||c||}(< c, u > + b)$$

Entonces el problema de hallar el hiperplano H consiste en resolver el siguiente problema lineal:

$$\begin{cases} \max\ M \\ s.a. \\ \| x \| = 1, \\ y_i(< a_i, x > + b) \geq M,\ i = 1, ..., m \end{cases} \qquad (1.1)$$

que es equivalente al siguiente problema de programación cuadrática

$$\begin{cases} \min \frac{1}{2} \| x \|^2 \\ s.a \\ y_i(< x, a_i > + b) \geq 1, \ i = 1,2,...,m \end{cases} \qquad (1.2)$$

La solución óptima (x, b) de (1.2) permite determinar los vectores soporte a_i como aquellos que satisfacen la relación:$< x, a_i > + b = \pm 1$; de aquí resultan los dos hiperplanos frontera:

H(+): < x, a_i > + b-1=0 y H(–): < x, a_i > + b+1=0

cuyos coeficientes son "x" y los términos independientes (b-1) y (b+1) respectivamente.

Denotaremos con F(-) y F(+) a los vectores soporte que están en los hiperplanos H(–) y H(+) respectivamente.

Las clases A(+) y A(–) se determinan mediante la función:

res = signo (< x, a_i > + b)

El punto a_i estará en la clase A(+) si res = +1 y estará en la clase A(–) si res = -1.

Como se mencionó en la sección 3.1.2 en la Base de datos, la variable Días de atraso de la última cuota pagada (V8) es la variable más realista, porque es la respuesta del sistema. La variable de aceptación o rechazo del crédito (V27) es artificial, porque depende de V8. Teniendo en cuenta la dependencia entre estas dos variables, se harán dos pruebas con la Máquina con Soporte Vectorial, una con la Base de datos completa, y otra con la variable V8 en reemplazo de la variable V27.

3.4.1. Prueba con la base de datos completa (27 variables)

La base de datos BD con 15569 registros y 26 variables, definen la matriz D de 15569x26 y la variable V27 define la matriz de las categorías F, de 15569x1.

a) Determinación de los Vectores Soporte

Para desarrollar la Máquina con Soporte Vectorial con núcleo lineal con datos de entrada D y matriz de categorías F, se ejecuta la función fitcsvm() con MatLab, y se obtienen una serie de archivos de la clase compact.

b) Los Vectores Soporte

Los vectores soporte se obtienen al ejecutar la sentencia indicada. Se obtiene la matriz 41x26 de vectores soporte de las clases A(+) y A(-).

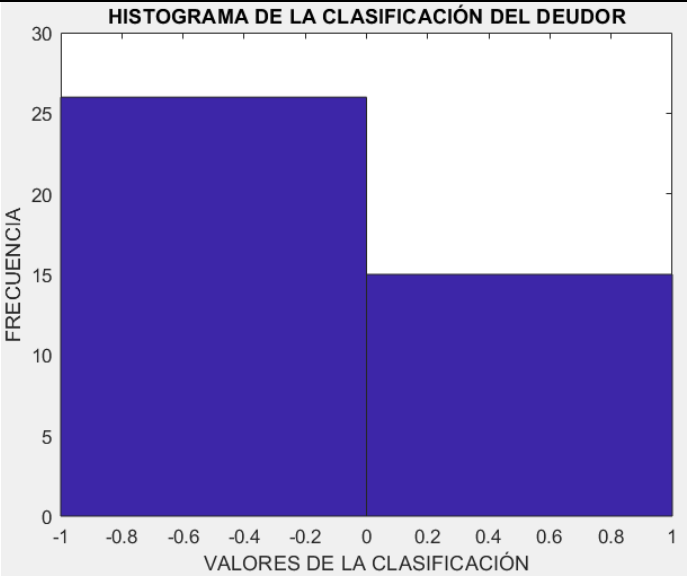
Luego de asociar a cada vector soporte, su categoría, se obtiene una matriz de 41x1; los valores “+1” y “-1” indican la clase.

Finalmente, se tiene en una sola matriz los vectores soporte y el vector de categorías.

Así se tiene una matriz de 41x27 donde las primeras 26 columnas son los vectores de entrenamiento (datos de entrada) y la columna 27 indica a que clase corresponden estos vectores A(+) o A(-).

Se sabe que hay un total de 313 créditos rechazados y 15256 créditos aceptados. Los vectores de soporte que se encontraron corresponden a 26 rechazados y 15 aceptados. En la Figura 36 se muestra la distribución de frecuencia.

Figura 36. **Distribución de frecuencia de los vectores soporte**

CENTROS	FREC	PROBAB	N° DE VECTORES SOPORTE
-1.00	26.00	0.63	
1.00	15.00	0.37	

Fuente: Obtenido con Matlab R2018a

3.4.2. Prueba con la base de datos (26 variables)

La base de datos BD con 15569 registros y 25 variables (eliminada V8), definen la matriz D de 15569x25 y la variable V27 define la matriz de las categorías F, de 15569x1.

a) Determinación de los vectores de soporte

Para desarrollar la Máquina con Soporte Vectorial con núcleo lineal con datos de entrada D y matriz de categorías F, se ejecuta la función fitcsvm () en MatLab, y se obtienen una serie de archivos de la clase compact.

b) Los Vectores Soporte

Los vectores soporte se obtienen al ejecutar la sentencia indicada. Se obtiene la matriz 28x25 de vectores soporte de las clases A(+) y A(-).

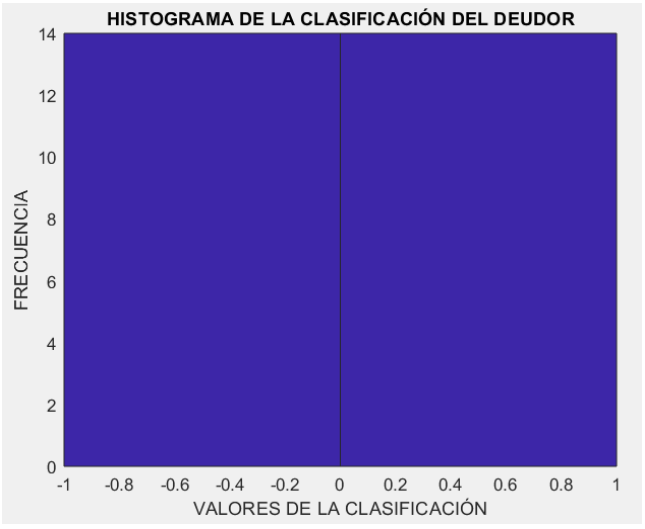
Luego de asociar a cada vector soporte, su categoría, se obtiene una matriz de 28x1; los valores “+1” y “-1” indican la clase.

Finalmente, se tiene en una sola matriz los vectores soporte y el vector de categorías.

Así se tiene una matriz de 28x26 donde las primeras 25 columnas son los vectores de entrenamiento (datos de entrada) y la columna 26 indica a que clase corresponden estos vectores A(+) o A(-).

Se sabe que hay un total de 313 créditos rechazados y 15256 créditos aceptados. Los vectores de soporte que se encontraron corresponden a 14 rechazados y 14 aceptados. En la Figura 37 se muestra la distribución de frecuencia de los vectores soporte.

Figura 37. **Distribución de frecuencia de los vectores soporte.**

CENTROS	FREC	PROBAB	N° DE VECTORES SOPORTE
-1.00	14.00	0.50	
1.00	14.00	0.50	

Fuente: Obtenido con Matlab R2018

CAPÍTULO IV: ANÁLISIS Y DISCUSIÓN DE RESULTADOS

En este capítulo se analiza la Base de datos y se discuten los resultados obtenidos con las pruebas usando técnicas de aprendizaje de máquina. Con las Redes Neuronales Artificiales Backpropagation (RNA) se ha hecho la predicción del comportamiento crediticio de los prestatarios, con las Redes Self-Organizing-Maps (RNA-SOM) se han agrupado los prestatarios en clústeres, y con la Máquina con Soporte Vectorial se han separado los registros de la base de datos en dos grupos, aceptados y rechazados.

4.1. ANÁLISIS DE LA BASE DE DATOS

a) Variables

En la sección 3.1.2 se han calculado las medidas estadísticas de las variables y el coeficiente de correlación de todas las variables con la variable Aceptación o rechazo del crédito (V27), los resultados se vuelven a presentar en la Tabla 8.

Tabla 8. Medidas estadísticas y coeficiente de correlación de la Base de datos

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.03	1	1	-0.01
V2	300	297770	10957.9	5879.45	1000	0.01
V3	16.24	297770	10475.32	5078.65	1000	0.01
V4	8	13	1.1	11.07	12	-0.03
V5	0	4	0.59	0.15	0	0.43
V6	0	4	0.55	0.13	0	0.45
V7	-419	268	37.64	-14.98	-15	0.26
V8	**0**	**162**	**8.32**	**2.88**	**0**	**0.78**
V9	-125	88	4.6	0.02	0	0.21
V10	0.83	48000	756.33	136.1	20	0.13

V11	0	297770	10480.37	5034.86	1000	0
V12	0	26527	453.83	38.05	0	0.22
V13	0	29822	350.02	5.75	0	0.04
V14	0	19652.81	498.09	164.71	0	0.05
V15	0	5903	94.62	9.03	0	0.19
V16	40546	40847	84.95	40711.96	40751	-0.14
V17	1	5	0.35	2.94	3	0.03
V18	0	360	21.54	6.35	0	0.03
V19	40714	48131	438.13	41198.31	41001	-0.01
V20	40579	41209	37.64	40862.16	40849	-0.26
V21	1	365	35.29	35.84	30	-0.02
V22	1	240	14.3	15.65	12	0.01
V23	0	37	2.77	3.47	0	0.03
V24	1	1	0	1	1	NaN
V25	1	44	13.38	18.98	1	-0.01
V26	12.01	293.79	18.25	51.09	58.27	0.04
V27	0	1	0.14	0.02	0	1

Fuente: Elaboración propia

Del análisis de las medidas estadísticas se observa lo siguiente:

(i) Con respecto a las correlaciones de la variable de aceptación o rechazo del crédito (V27) con las 26 primeras variables de la Base de datos, la variable V8, que indica los Días de atraso de la última cuota pagada, es la variable que tiene la mayor correlación (0.78).

(ii) Las siguientes variables con mayor correlación son: Clasificación del deudor (V5) con valor 0.43 y Clasificación del deudor sin considerar alineamiento con el sistema (V6) con valor 0.45. La variable que tiene correlación cero es Saldo capital vigente de la operación (V11).

(iii) La variable Indicador de Rescate Financiero Agropecuario (V24) tiene una correlación indeterminada, esto se explica porque todos los registros de la Base de datos tienen el valor constante “1”.

(iv) Tal como se mencionó en la sección 3.1.2, la variable V8 es la salida de respuesta real del sistema usado por la entidad de micro-finanzas, porque está dando una variabilidad de respuesta. Por tanto, se puede afirmar que la variable V27 (con valores 0 y 1) es artificial, porque anula un estudio inteligente de la Base de datos, debido a que depende de V8. Así, cuando V8 varía de 0 a 30 días, V27 vale "0", esto equivale a comprimir los valores de 0 a 30 y asignarle el valor 0, por otro lado, cuando V8 varía de 31 a más días, se le asigna el valor "1". Por esta razón es que se han hecho pruebas incluyendo la variable V8 en un caso y excluyéndolo en otro caso, tanto con las Redes Neuronales Artificiales como también con las Máquinas con Soporte Vectorial.

b) Análisis de las sub-bases de datos BDR y BDA

En la base de datos se observa que los valores de la variable V8 que están en el intervalo [0, 30] están asociados con el valor "0" de la variable V27 y, que los registros que están en el intervalo [31, 162] están asociados con el valor "1". Basados en estos resultados, se separan las 26 primeras variables de la base de datos en dos sub-bases: Los clientes con crédito aceptado (V27=0), que se denominará BDA con un total de 15,256 registros, y los correspondientes a los clientes con crédito rechazado (V27=1), que se denominará BDR que tiene un total de 313 registros.

Con la finalidad de analizar en detalle el comportamiento de las variables, se calculan los mínimos, máximos, la desviación estándar, la media y la moda de las variables de cada una de estas dos sub-bases. Las Tablas 9 y 10 muestran los estadísticos señalados para las 26 variables de las dos sub-bases BDR y BDA.

Tabla 9. Medidas estadísticas de la sub-base de datos de créditos rechazados

Variable	Min	Max	STD		Media	Moda
V1	1	1	0		1	1
V2	500	45418.18	12141.35		6879.11	1000
V3	93.11	44387.21	11697.96		6008.63	645.19
V4	9	12	1.17		10.8	12
V5	0	4	1.19		1.92	1
V6	0	4	1.19		1.87	1
V7	-133	230	49.75		53.62	16
V8	31	162	18.31		48.21	32
V9	-2	88	10.83		6.72	0
V10	7.22	29821.89	2124.62		842.89	38.58
V11	0	44387.21	11735.79		5174.49	0
V12	0	26527	2197.91		738.86	0
V13	0	29822	1685.64		95.28	0
V14	0	3420.78	845.31		333.61	0
V15	0	3255	306.95		137.83	0
V16	40548	40777	54.02		40629.67	40633
V17	3	5	0.11		3.01	3
V18	0	120	30.99		10.21	0
V19	40739	42598	442.83		41157.6	40941
V20	40617	40980	49.75		40793.38	40831
V21	7	180	8.58		30.41	30
V22	2	60	13.79		17.04	12
V23	0	11	2.07		3.96	3
V24	1	1	0		1	1
V25	1	43	12.58		18.07	15
V26	18.02	79.59	17.63		56.04	58.27

Fuente: Elaboración propia

Tabla 10. Medidas estadísticas de la sub-base de datos de créditos Aceptados

Variable	Min	Max	STD	Media	Moda
V1	1	2	0.04	1	1
V2	300	297770	10931.76	5858.94	1000
V3	16.24	297770	10448.3	5059.57	1000
V4	8	13	1.09	11.07	12
V5	0	4	0.5	0.11	0
V6	0	4	0.47	0.09	0
V7	-419	268	36.01	-16.39	-15
V8	0	30	4.56	1.95	0
V9	-125	62	4.28	-0.11	0
V10	0.83	48000	693.54	121.6	20
V11	0	297770	10453.44	5032	1000
V12	0	25104	317.97	23.67	0
V13	0	20000	258.36	3.91	0
V14	0	19652.81	487.82	161.24	0
V15	0	5903	82.84	6.39	0
V16	40546	40847	84.63	40713.65	40847
V17	1	5	0.35	2.94	3
V18	0	360	21.3	6.27	0
V19	40714	48131	438.01	41199.15	41001
V20	40579	41209	36.01	40863.57	40849
V21	1	365	35.62	35.95	30
V22	1	240	14.31	15.62	12
V23	0	37	2.78	3.46	0
V24	1	1	0	1	1
V25	1	44	13.4	18.99	1
V26	12.01	293.79	18.25	50.99	58.27

Fuente: Elaboración propia

Del estudio de las Tablas 9 y 10 se tienen las siguientes observaciones:

(i) En la Tabla 9, de los créditos rechazados, la variable V8 varía en el intervalo [31, 162], y en la Tabla 10, la variable V8 para los créditos aceptados BDA, varía en el intervalo [0, 30], estos rangos son disjuntos.

En consecuencia, el criterio de calificación de los clientes en rechazados o aceptados se define mediante estos dos rangos. Es decir, los clientes son:

Rechazados si V8 >= 31

Aceptados si 0<=V8<=30

Esta situación indica que las 25 variables diferentes de la variable V8 no intervienen en la calificación de los clientes en Aceptados o Rechazados, lo cual demuestra que el sistema que emplea la entidad de microcréditos no está haciendo un uso inteligente de las otras variables que ayudarían a tomar una decisión más acertada para el rechazo o la aceptación de un cliente.

(i) Las sub-bases de datos de los clientes aceptados BDA y de los clientes rechazados BDR tienen 15,236 y 313 registros respectivamente.

(ii) Desde el punto de vista práctico, son importantes las variables Monto del crédito otorgado (V2) y Número de cuotas programadas (V22).

En el grupo de créditos rechazados:

Los valores de la variable V2 van de 500 a 45,418 soles con moda 1000 soles, media de 6,879 soles y desviación estándar 12,141 soles. Esta última medida estadística indica que hay una alta variabilidad.

Los valores de la variable V22 van de 2 a 60 meses, con moda 12 meses, media de 17.04 meses y desviación estándar 13.79

En el grupo de créditos aceptados:

Los valores de la variable V2 van de 300 a 297,770 soles con desviación estándar 10,932 soles, media de 5,859 soles y moda de 1000 soles. Esto

último indica una menor variabilidad con respecto a los créditos rechazados, y también indica que los créditos de monto alto son pocos en número, pero son clientes que han cumplido. Se podría haber pensado que los rechazados correspondían a los clientes de montos altos, pero esta medida indica que no es así.

Los valores de la variable V22 varían de 1 a 240 meses, con desviación estándar 14.31, media 15.62 y moda 12 meses.

4.2. RESULTADOS DE LA INVESTIGACIÓN

4.2.1. Redes Neuronales Artificiales Backpropagation (RNA)

Con las RNA Backpropagation se ha obtenido la predicción del comportamiento crediticio de los prestatarios mediante el valor de R que indica la precisión de la predicción. En cada caso se han probado diferentes arquitecturas de la red neuronal, y se ha determinado el valor de la precisión R para las etapas de entrenamiento, validación, prueba y la precisión global.

a) Base de datos completa (27 variables)

a.1) Prueba con una RNA de 4 capas con 14, 10, 8 y 1 neuronas

Etapa	*Entrenamiento*	*Validación*	*Prueba*	*Global*
R	0.99209	0.96968	0.95715	0.97682

a.2) Prueba con una RNA de 4 capas con 20, 14, 8 y 1 neuronas

Etapa	*Entrenamiento*	*Validación*	*Prueba*	*Global*
R	0.99181	0.95673	0.93735	0.96794

a.3) Prueba con una RNA de 3 capas con 20, 10 y 1 neuronas

Etapa	*Entrenamiento*	*Validación*	*Prueba*	*Global*
R	0.95303	0.91223	0.89767	0.92657

a.4) Prueba con una RNA de 5 capas con 20, 18, 14, 10 y 1 neuronas

Etapa	*Entrenamiento*	*Validación*	*Prueba*	*Global*
R	0.99927	0.94617	0.91187	0.96227

La Tabla 11 muestra un consolidado de la precisión global obtenida para las diversas arquitecturas de la RNA con la Base de datos completa. Se aprecia que la mayor precisión global es 0.97682 que se obtiene con una RNA Backpropagation de 4 capas, con 14, 10, 8 y 1 neuronas respectivamente en cada capa.

Tabla 11. Precisión global de las diferentes arquitecturas con BD completa

N° de capas	Neuronas	R Global
4	[14, 10, 8, 1]	**0.97682**
4	[20, 14, 8 ,1]	0.96794
3	[20, 10, 1]	0.92657
5	[20, 18, 14, 10, 1]	0.96227

Fuente: Elaboración propia

b) Base de datos con 26 variables (V8 reemplaza a V27)

b.1) Prueba con una RNA de 4 capas con 14, 10, 8 y 1 neuronas

Etapa	*Entrenamiento*	*Validación*	*Prueba*	*Global*
R	0.79728	0.74039	0.7267	0.76192

b.2) Prueba con una RNA de 4 capas con 20, 14, 8 y 1 neuronas

Etapa	*Entrenamiento*	*Validación*	*Prueba*	*Global*
R	0.82129	0.75122	0.72003	0.77389

b.3) Prueba con una RNA de tres capas con 20, 10 y 1 neuronas

Etapa	*Entrenamiento*	*Validación*	*Prueba*	*Global*
R	0.68305	0.63127	0.62338	0.65224

b.4) Prueba con una RNA de 5 capas con 20, 18, 14, 10 y 1 neuronas

Etapa	*Entrenamiento*	*Validación*	*Prueba*	*Global*
R	0.80725	0.73132	0.70996	0.7596

La Tabla 12 muestra un consolidado de la precisión global obtenida para las diversas arquitecturas de la RNA con la Base de datos completa. Se aprecia que la mayor precisión global es 0.77389 que se obtiene con una RNA Backpropagation de 4 capas, con 20, 14, 8 y 1 neuronas respectivamente en cada capa.

Tabla 12. Precisión global de las diferentes arquitecturas, BD con 26 variables (V8 reemplaza a V27)

N° de capas	Neuronas	R Global
4	[14, 10, 8, 1]	0.76192
4	[20, 14, 8 ,1]	**0.77389**
3	[20, 10, 1]	0.65224
5	[20, 18, 14, 10, 1]	0.7596

Fuente: Elaboración propia

c) Base de datos con 26 variables (eliminada V8)

c.1) Prueba con una RNA de 4 capas con 14, 10, 8 y 1 neuronas

Etapa	*Entrenamiento*	*Validación*	*Prueba*	*Global*
R	0.7871	0.64229	0.48428	0.66897

c.2) Prueba con una RNA de 4 capas con 20, 14, 8 y 1 neuronas.

Etapa	*Entrenamiento*	*Validación*	*Prueba*	*Global*
R	0.73685	0.62707	0.56765	0.66304

c.3) Prueba con una RNA de 3 capas con 20, 10 y 1 neuronas

Etapa	*Entrenamiento*	*Validación*	*Prueba*	*Global*
R	0.5875	0.55382	0.53935	0.56586

c.4) Prueba con una RNA de 5 capas con 20, 18, 14, 10 y 1 neuronas

Etapa	*Entrenamiento*	*Validación*	*Prueba*	*Global*
R	0.69511	0.56607	0.52867	0.61637

La Tabla 13 muestra un consolidado de la precisión global obtenida para las diversas arquitecturas de la RNA con la Base de datos completa. Se aprecia que la mayor precisión global es 0.66897 que se obtiene con una RNA Backpropagation de 4 capas, con 14, 10, 8 y 1 neuronas respectivamente en cada capa.

Tabla 13. Precisión global de las diferentes arquitecturas de RNA con 26 variables (eliminada V8)

N° de capas	Neuronas	R Global
4	[14, 10, 8, 1]	**0.66897**
4	[20, 14, 8 ,1]	0.66304
3	[20, 10, 1]	0.56586
5	[20, 18, 14, 10, 1]	0.61637

Fuente: Elaboración propia

La Tabla 14 muestra un consolidado de las pruebas que se han hecho con la Base de datos con diferentes dimensiones.

Tabla 14. Resultados de todas las pruebas con RNA

Base de datos	R Global	Arquitectura	Observación
BD1 (27 variables)	0.97682	[14, 10, 8, 1]	Base de datos completa, con 26 variables independientes y la variable V27 de aceptación o rechazo del crédito.
BD2 (26 variables)	0.77389	[20, 14, 8 ,1]	Base de datos con 26 variables, en la que la variable V8 reemplaza a la variable V27.
BD3 (26 variables)	0.66897	[14, 10, 8, 1]	Base de datos con 26 variables, eliminada la variable V8.

Fuente: Elaboración propia

d) Conclusiones

De las pruebas que se han hecho con las RNA Backpropagation con variaciones en la base de datos y diferentes arquitecturas, se tienen las siguientes conclusiones:

(i) La mayor precisión es 0.97682 y se ha obtenido con la base de datos completa. Esta alta precisión de la red neuronal se explica porque la variable V27 depende directamente de la variable V8.

(ii) Al reemplazar la variable V27 por la variable V8, la precisión obtenida por la red neuronal disminuye a 0.77389, pero sigue siendo una buena precisión. Se puede afirmar que la variable V8 es una variable de respuesta que está oculta, y es la más realista.

(iii) Al eliminar la variable V8 de la base datos, la precisión de la red neuronal baja a 0.66897, con lo cual se corrobora que la variable V8 es la más realista, y que la variable V27 es artificial.

(iv) Se comprueba que un mayor número de neuronas en la capa oculta no implica una mejora en la precisión de la RNA Backpropagation.

4.2.2. Redes Self-Organizing-Maps (RNA-SOM)

Con las RNA-SOM se agrupan los prestatarios en clústeres.

(I) Red Neuronal SOM: 2 neuronas, Gridtop, Dist

En esta prueba se tienen dos clústeres C1 y C2 con 1,033 y 14,536 registros respectivamente, a continuación, se analizarán las características de los registros de cada clúster.

Variables influyentes en la formación de los clústeres

En la Tabla 15 se presenta las estadísticas de las 27 variables y el coeficiente de correlación con el clúster. Para identificar las variables que más han influido en la formación de los clústeres se determina el coeficiente de correlación de las 27 variables con el clúster.

Tabla 15. Estadísticas y CC de col 28 (clústeres) con las 27 variables

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.03	1	1	-0.1
V2	**300**	**297770**	**10957.9**	**5879.45**	**1000**	**-0.84**
V3	**16.24**	**297770**	**10475.32**	**5078.65**	**1000**	**-0.84**
V4	8	13	1.1	11.07	12	0.35
V5	0	4	0.59	0.15	0	0.03
V6	0	4	0.55	0.13	0	0.04
V7	-419	268	37.64	-14.98	-15	0.08
V8	0	162	8.32	2.88	0	-0.02
V9	-125	88	4.6	0.02	0	0
V10	0.83	48000	756.33	136.1	20	-0.24
V11	**0**	**297770**	**10480.37**	**5034.86**	**1000**	**-0.84**
V12	0	26527	453.83	38.05	0	0.02
V13	0	29822	350.02	5.75	0	0

V14	**0**	**19652.81**	**498.09**	**164.71**	**0**	**-0.69**
V15	0	5903	94.62	9.03	0	0.01
V16	40546	40847	84.95	40711.96	40751	0.06
V17	1	5	0.35	2.94	3	0.02
V18	0	360	21.54	6.35	0	-0.6
V19	40714	48131	438.13	41198.31	41001	-0.57
V20	40579	41209	37.64	40862.16	40849	-0.08
V21	1	365	35.29	35.84	30	-0.03
V22	1	240	14.3	15.65	12	-0.55
V23	0	37	2.77	3.47	0	0.07
V24	1	1	0	1	1	NaN
V25	1	44	13.38	18.98	1	-0.01
V26	12.01	293.79	18.25	51.09	58.27	0.39
V27	-1	1	0.28	0.96	1	0.02
Clústeres	1	2	0.25	1.93	2	1

Fuente: Elaboración propia

La Tabla 16 presenta un resumen de las variables que tienen la correlación más alta.

Tabla 16. Variables con la correlación más alta

Variable	Coeficiente de correlación	Descripción
V2	-0.84	Monto del crédito otorgado.
V3	-0.84	Saldo capital de la deuda.
V11	-0.84	Saldo de capital vigente de la operación.
V14	-0.69	Rendimientos devengados de la operación.

Fuente: Elaboración propia

Con respecto a las estadísticas y al coeficiente de correlación, se comenta lo siguiente:

(i) Se observa, por el valor de las correlaciones, las variables que más han influido en la formación de los clústeres son el monto del crédito otorgado (V2), el saldo capital de la deuda (V3), el saldo capital vigente de la operación (V11), y los rendimientos devengados de la operación (V14).

La alta correlación que tienen estas variables explica por qué son las que han determinado la formación de los clústeres.

(ii) Estas variables presentan valores grandes comparado con las otras variables, al observar la media de las otras variables, los valores de estas variables son altas, en los rangos, los valores máximos son altos. Lo que marca la contribución de una variable en el proceso de agrupamiento es la media, y este valor es alto en estas variables.

(iii) La contribución de estas variables en la formación de los clústeres, se explica porque se están usando distancias euclidianas entre los datos del cliente con el centro de las neuronas y la topología es de las mallas rectangulares, entonces los valores numéricamente grandes están influyendo en la formación de los clústeres.

(iv) En el campo de las aplicaciones, la red neuronal está privilegiando a los clientes que tienen préstamos más grandes, es decir ha clasificado en función del monto del préstamo.

(v) También es necesario comentar cómo ha sido la participación de la variable de aceptación o rechazo de la solicitud de crédito (V27) en la formación de los clústeres. El coeficiente de correlación de V27 de valor 0.024 indica que su participación ha sido muy baja.

Estudio del clúster C1

a) Correlaciones en el clúster C1

La Tabla 17 muestra las medidas estadísticas de las variables y su correlación con la variable V27 en el clúster C1.

Tabla 17. Estadísticas del clúster C1 (CC con V27)

VARIABLE	MIN	MAX	STD	MEDIA	MODA	CCf
V1	1	2	0.12	1.01	1	0.02
V2	20,880	297,770	17,293.06	40,271.76	40,000	0.01
V3	17,234.95	297,770	17,142.54	38,057.59	41,625.02	0.01
V4	8	13	1.08	9.63	9	0.09
V5	0	3	0.3	0.07	0	-0.44
V6	0	1	0.23	0.06	0	-0.56
V7	-419	30	39.19	-26.38	-15	-0.2
V8	**0**	**60**	**8.79**	**3.55**	**0**	**-0.78**
V9	-18	31	1.65	-0.03	0	-0.11
V10	229.68	48000	2241.95	809.28	600	-0.08
V11	17234.95	297770	17142.54	38057.59	41625.02	0.01
V12	0	0	0	0	0	NaN
V13	0	0	0	0	0	NaN
V14	0	19652.81	1220.78	1456.49	1204.62	-0.17
V15	0	2791	129.05	7.22	0	0.01
V16	40547	40847	87	40694.05	40786	0.16
V17	1	3	0.41	2.91	3	-0.04
V18	0	120	56.33	54.94	120	-0.09
V19	40852	48131	797.04	42128.66	42700	-0.03
V20	40817	41209	38.66	40873.44	40862	0.2
V21	30	365	45.42	39.65	30	0.04
V22	1	240	26.45	44.97	60	-0.05
V23	0	9	2.29	2.74	0	-0.04
V24	1	1	0	1	1	NaN
V25	1	43	10.93	19.33	15	0.02
V26	12.01	49.36	7.06	24.14	18.02	0.02

V27	-1	1	0.36	0.93	1	1
Clúster C1	1	1	0	1	1	NaN

Fuente: Elaboración propia

Con respecto a los resultados mostrados en la Tabla 17 se observa lo siguiente:

(i) El clúster ha seleccionado clientes cuyo monto del crédito otorgado (V2), mayormente se concentran en alrededor de la media de 40,272 soles. El máximo crédito es de 297,770 soles, hay un reducido número de clientes que tienen ese valor máximo, esto se explica por el valor de la media. El monto mínimo de crédito en este clúster es de 20,880 soles, que es el otro extremo, y también lo tienen un reducido número de clientes. La desviación estándar nos permite afirmar que hay una franja alrededor de la media en donde se encuentran el monto de préstamo de la mayoría de los clientes, y que estos oscilan entre los valores 22,979 soles y 57,565 soles. Esta variable ha tenido una participación significativa en la formación de los clústeres, esto se explica por el valor de su alta correlación (-0.84), en cambio, su correlación con la variable V27 es muy baja, lo que indica que V2 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(ii) Con respecto al saldo capital de la deuda (V3), el clúster ha seleccionado clientes cuyo saldo capital se concentran en alrededor de la media de 38,058 soles. El monto máximo de saldo capital es de 297,770 soles, que lo tienen un reducido número de clientes, esto se explica por el valor de la media. El monto mínimo de saldo capital 17,235 soles, y también lo tienen un reducido número de clientes. La desviación estándar nos permite afirmar que hay una franja alrededor de la media en donde se encuentran el saldo capital de la mayoría de los clientes, y que estos oscilan entre los valores 20,915 soles y 55,200 soles. Esta variable también ha sido tomada en cuenta para la formación de los clústeres, esto

se explica por el valor de su alta correlación (–0.84). En cambio, su correlación con la variable V27 es muy baja, lo que indica que V3 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iii) La variable saldo de capital vigente de la operación (V11) varía en el rango de 17,235 a 297,770 soles, su media es 38,058 soles. Hay un reducido número de clientes con un saldo de capital vigente de la operación con valores iguales o cercanos a los extremos, esto se explica por el valor de la media. Teniendo en cuenta la desviación estándar y la media, se puede afirmar que la mayoría de los clientes tienen un saldo de capital vigente de la operación que fluctúa entre los valores de 20,915 y 55,200 soles. La variable V11 también tiene un coeficiente de correlación (-0.84), lo que explica porque se ha tomado en cuenta en la formación de los clústeres. También es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que V11 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iv) La variable rendimientos devengados de la operación (V14) tiene valores que fluctúan desde un mínimo de cero hasta 19,653 soles. Teniendo en cuenta que la media es de 1456 soles, se puede afirmar que hay un reducido número de clientes con valores altos de esta variable, la misma afirmación se puede hacer para el número de clientes que tienen valor cero para esta variable. Los rendimientos devengados de la operación de la mayoría de los clientes fluctúan entre los valores de 236 y 2,677 soles. La variable V14 tiene un coeficiente de correlación significativamente alto (-0.69), lo que explica porque se ha tomado en cuenta en la formación de los clústeres. Es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que V14 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(v) La variable días de atraso de la última cuota pagada (V8) varía en el rango de 0 a 60 días. Su media de 3.55 indica que la mayoría de los clientes ha pagado con un retraso de hasta 30 días, muy pocos son los clientes que han pagado con un retraso igual o cercano a los 60 días. Esta variable tiene una correlación baja en la formación de los clústeres, pero dentro del clúster C1, es la variable que tiene la más alta correlación (-0.78) con la variable de aceptación o rechazo de la solicitud de crédito (V27).

b) Frecuencia de las variables Aceptación o rechazo de la solicitud de crédito (V27) y Días de atraso de la última cuota pagada (V8) en el clúster C1

Tabla 18. Frecuencia de las variables V27 y V8

Frecuencia de V27			Frecuencia de V8		
-1 v 1	FREC	PROBAB	[0 , 60]	FREC	PROBAB
-1.00	34.00	0.03	0	794.00	0.77
1.00	999.00	0.97	5.00	98.00	0.09
			10.00	39.00	0.04
			15.00	34.00	0.03
			20.00	15.00	0.01
			25.00	11.00	0.01
			30.00	14.00	0.01
			35.00	10.00	0.01
			40.00	4.00	0.00
			45.00	7.00	0.01
			50.00	2.00	0.00
			55.00	4.00	0.00
			60.00	1.00	0.00

Fuente: Elaboración propia

c) Frecuencia de las variables Clasificación del deudor (V5) y Clasificación del deudor sin considerar alineamiento del sistema (V6) en el clúster C1

Tabla 19. Frecuencia de las variables V5 y V6

Frecuencia de V5			Frecuencia de V6		
[0, 3]	**FREC**	**PROBAB**	**[0, 1]**	**FREC**	**PROBAB**
0	971.00	0.94	0	974.00	0.94
1.00	56.00	0.05	1.00	59.00	0.06
2.00	2.00	0.00			
3.00	4.00	0.00			

Fuente: Elaboración propia

Comentarios sobre la frecuencia de las variables V5 y V6 en el clúster C1:

Las variables V5 y V6, después de la variable V8, dentro del clúster C1 son las variables con la correlación más alta con respecto a la variable V27.

(i) Se observa que en el clúster C1, considerando el factor de la SBS para la clasificación de los deudores, el 94% de los clientes ha sido clasificado como un deudor normal, el 5% de los clientes ha sido clasificados como un cliente con potencial pérdida, y el resto ha sido clasificado como deficiente o dudoso.

(ii) Con respecto a la clasificación propia de la entidad financiera sin considerar el factor de la SBS, en este clúster el 94% de los clientes ha sido clasificado como un deudor normal, y el 6% ha sido clasificado como un cliente con potencial pérdida.

d) Frecuencia de las variables Monto del crédito otorgado (V2) y Saldo capital de la deuda (V3) en el clúster C1

Tabla 20. Frecuencia de las variables V2 y V3

Frecuencia de V2			Frecuencia de V3		
[20880, 297770]	FREC	PROBAB	[17234.95,297770]	FREC	PROBAB
20880.00	271.00	0.26	17234.95	240.00	0.23
43954.17	712.00	0.69	40612.87	740.00	0.72
67028.33	20.00	0.02	63990.79	25.00	0.02
90102.50	17.00	0.02	87368.71	15.00	0.01
113176.67	8.00	0.01	110746.63	9.00	0.01
136250.83	2.00	0.00	134124.55	1.00	0.00
159325.00	1.00	0.00	157502.48	1.00	0.00
182399.17	0	0	180880.40	0	0
205473.33	0	0	204258.32	0	0
228547.50	0	0	227636.24	0	0
251621.67	0	0	251014.16	0	0
274695.83	1.00	0.00	274392.08	1.00	0.00
297770.00	1.00	0.00	297770.00	1.00	0.00

Fuente: Elaboración propia

e) Frecuencia de las variables Saldo capital vigente de la operación (V11) y Rendimientos devengados de la operación (V14) en el clúster C1

Tabla 21. Frecuencia de las variables V11 y V14

Frecuencia de V11			Frecuencia de V14		
[17234.95, 297770]	FREC	PROBAB	[0, 19652.81]	FREC	PROBAB
17234.95	240.00	0.23	0	425.00	0.41
40612.87	740.00	0.72	1637.73	385.00	0.37
63990.79	25.00	0.02	3275.47	212.00	0.21
87368.71	15.00	0.01	4913.20	7.00	0.01
110746.63	9.00	0.01	6550.94	2.00	0.00
134124.55	1.00	0.00	8188.67	0	0

157502.48	1.00	0.00	9826.41	1.00	0.00
180880.40	0	0	11464.14	0	0
204258.32	0	0	13101.87	0	0
227636.24	0	0	14739.61	0	0
251014.16	0	0	16377.34	0	0
274392.08	1.00	0.00	18015.08	0	0
297770.00	1.00	0.00	19652.81	1.00	0.00

Fuente: Elaboración propia

f) Rentabilidad en el clúster C1

- Total créditos aceptados: 999
- Monto de crédito promedio: S/ 40,271.76
- Total monto de créditos otorgados: 40’231,488
- Plazo de pago promedio: 45 meses
- Intereses ganados: 27’238,245
- Rentabilidad en el período de pago: 67.7%
- Rentabilidad anual: 18.05%

g) Riesgo en el clúster C1

- Máximo días de atraso de la última cuota pagada: 60 días
- Promedio días de atraso de la última cuota pagada: 3.55 días
- Riesgo: 5.92 %

Estudio del clúster C2

a) Correlaciones en el clúster C2

La Tabla 22 muestra las medidas estadísticas de las variables en el clúster C2 y su correlación con la variable V27.

Tabla 22. Estadísticas del Clúster C2 (CC con V27)

Variable	Min	Max	STD	Media	Moda	CCf
V1	1	2	0.02	1	1	0
V2	300	35000	4164	3435.36	1000	0.02
V3	16.24	29821.89	3725.17	2735.01	1000	0.02
V4	9	13	1.02	11.17	12	0.02
V5	0	4	0.6	0.15	0	-0.44
V6	0	4	0.57	0.14	0	-0.46
V7	-360	268	37.4	-14.17	-2	-0.27
V8	**0**	**162**	**8.28**	**2.84**	**0**	**-0.78**
V9	-125	88	4.74	0.03	0	-0.22
V10	0.83	29821.89	470.45	88.27	20	-0.19
V11	0	20609.59	3710.81	2688.1	1000	0.06
V12	0	26527	469.57	40.75	0	-0.23
V13	0	29822	362.24	6.15	0	-0.04
V14	0	6871.68	181.35	72.9	0	0.01
V15	0	5903	91.69	9.16	0	-0.22
V16	40546	40847	84.66	40713.24	40847	0.14
V17	1	5	0.35	2.94	3	-0.03
V18	0	360	9.6	2.9	0	0.02
V19	40714	47893	307.58	41132.2	41001	0.05
V20	40579	41207	37.44	40861.36	40849	0.27
V21	1	365	34.44	35.57	30	0.02
V22	1	240	10.2	13.56	12	0.01
V23	0	37	2.8	3.53	0	-0.03
V24	1	1	0	1	1	NaN
V25	1	44	13.54	18.95	1	0.01
V26	12.01	293.79	17.26	53.01	58.27	-0.06
V27	-1	1	0.27	0.96	1	1
Clúster C2	2	2	0	2	2	NaN

Fuente: Elaboración propia

En la Tabla 22 se observa lo siguiente:

(i) El clúster C2 ha seleccionado clientes cuyo monto del crédito otorgado (V2) es relativamente bajo, que están en el rango de 300 hasta 35,000 soles, estos créditos se concentran alrededor de la media de 3,435 soles. El máximo crédito es de 35,000 soles, que lo tienen un reducido número de clientes, esto se explica por el valor de la media. El monto mínimo de crédito es de 300 soles, y también lo tienen un reducido número de clientes. Esta variable ha tenido una participación importante en la formación de los clústeres, esto se explica por el valor de su correlación alta (-0.84), en cambio, su correlación con la variable V27 es muy baja, lo que indica que V2 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(ii) La variable saldo capital de la deuda (V3), indica que, en el período de repago del préstamo, el saldo del capital varía en el rango de 16 hasta 29,822 soles, siendo el promedio de 2,735 soles. Hay un reducido número de clientes que tienen un saldo capital de la deuda igual a los valores máximos o mínimo, esto se explica por el valor de la media. La mayoría de los clientes tienen un saldo de capital cuyo monto se encuentra alrededor de la media. El alto grado de correlación de esta variable (-0.84) explica porque esta variable ha sido tomada en cuenta para la formación de los clústeres. En cambio, su correlación con la variable V27 es muy baja, lo que indica que V3 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iii) La variable saldo de capital vigente de la operación (V11) varía en el rango de 0 a 20,610 soles, su media es 2,688 soles. Es reducido el número de clientes con un saldo de capital vigente de la operación con valores iguales o cercanos al máximo o al mínimo, es decir, la mayoría de los clientes tienen un saldo de capital vigente de la operación alrededor de la media. La variable V11 tiene un coeficiente de correlación alto cuyo valores -0.84 lo que explica porque se ha tomado en cuenta en la

formación de los clústeres. También es necesario precisar que la correlación de esta variable con la variable V27 es muy baja, lo que indica que V11 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iv) La variable rendimientos devengados de la operación (V14) tiene valores que fluctúan desde un mínimo de 0 hasta 6,872 soles. Teniendo en cuenta que la media es de 73 soles, se puede afirmar que hay un reducido número de clientes con valores iguales o cercanos al mínimo o máximo, por lo que, para la mayoría de los clientes, el valor de esta variable se concentra alrededor de la media. La variable V14 tiene un coeficiente de correlación alto con valor -0.69 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. Es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que V14 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(v) La variable días de atraso de la última cuota pagada (V8) en el clúster C2 varía en el rango de 0 a 162 días. Su media de 2.84 días es ligeramente menor a la media de esta variable en el clúster C1. Esto indica que la mayoría de los clientes ha tenido un retraso de 3 días para pagar su cuota, solo algunos clientes se han pasado de los 30 días. Esta variable tiene una correlación baja en la formación de los clústeres, sin embargo, es la variable que tiene la mayor correlación con la variable de aceptación o rechazo de la solicitud de crédito.

b) Frecuencia de las variables Aceptación o rechazo de la solicitud de crédito (V27) y Días de atraso de la última cuota pagada (V8) en el clúster C2

Tabla 23. Frecuencia de las variables V27 y V8

Frecuencia de V27			Frecuencia de V8		
-1 v 1	**FREC**	**PROBAB**	**[0, 162]**	**FREC**	**PROBAB**
-1.00	279.00	0.02	0	12973.00	0.89
1.00	14257.00	0.98	13.50	1041.00	0.07
			27.00	285.00	0.02
			40.50	130.00	0.01
			54.00	63.00	0.00
			67.50	20.00	0.00
			81.00	13.00	0.00
			94.50	3.00	0.00
			108.00	3.00	0.00
			121.50	2.00	0.00
			135.00	2.00	0.00
			148.50	0	0
			162.00	1.00	0.00

Fuente: Elaboración propia

c) Frecuencia de las variables Clasificación del deudor (V5) y Clasificación del deudor sin considerar alineamiento del sistema (V6) en el clúster C2

Tabla 24. Frecuencia de las variables V5 y V6

Frecuencia de V5			Frecuencia de V6		
[0, 4]	**FREC**	**PROBAB**	**[0, 4]**	**FREC**	**PROBAB**
0	13399.00	0.92	0	13471.00	0.93
1.00	573.00	0.04	1.00	597.00	0.04
2.00	211.00	0.01	2.00	171.00	0.01
3.00	215.00	0.01	3.00	159.00	0.01
4.00	138.00	0.01	4.00	138.00	0.01

Fuente: Elaboración propia

Comentarios sobre la frecuencia de las variables V5 y V6 en el clúster C2:

Las variables V5 y V6, son las que tienen mayor con V27 después de la variable V8 dentro del clúster C2.

(i) Se observa que en el clúster C2, considerando el factor de la SBS para la clasificación de los deudores, el 92% de los clientes ha sido clasificado como un deudor normal, el 4% de los clientes ha sido clasificados como un cliente con potencial pérdida, el resto ha sido clasificado como deficiente, dudoso o pérdida.

(ii) Con respecto a la clasificación propia de la entidad financiera sin considerar el factor de la SBS, en este clúster el 93% de los clientes ha sido clasificado como un deudor normal, el 4% ha sido clasificado como un cliente con potencial pérdida, el resto ha sido clasificado como deficiente, dudoso o pérdida.

d) Frecuencia de las variables Monto del crédito otorgado (V2) y Saldo capital de la deuda (V3) en el clúster C2

Tabla 25. Frecuencia de las variables V2 y V3

Frecuencia de V2			Frecuencia de V3		
[300, 35000]	**FREC**	**PROBAB**	**[16.24, 29821.89]**	**FREC**	**PROBAB**
300.00	6652.00	0.46	16.24	7030.00	0.48
3191.67	4697.00	0.32	2500.04	4628.00	0.32
6083.33	1464.00	0.10	4983.85	1224.00	0.08
8975.00	791.00	0.05	7467.65	508.00	0.03
11866.67	184.00	0.01	9951.46	410.00	0.03
14758.33	358.00	0.02	12435.26	200.00	0.01
17650.00	64.00	0.00	14919.06	255.00	0.02
20541.67	295.00	0.02	17402.87	102.00	0.01
23433.33	10.00	0.00	19886.67	176.00	0.01

26325.00	8.00	0.00	22370.48	0	0
29216.67	11.00	0.00	24854.28	1.00	0.00
32108.33	1.00	0.00	27338.09	1.00	0.00
35000.00	1.00	0.00	29821.89	1.00	0.00

Fuente: Elaboración propia

e) Frecuencia de las variables Saldo capital vigente de la operación (V11) y Rendimientos devengados de la operación (V14) en el clúster C2

Tabla 26. Frecuencia de las variables V11 y V14

Frecuencia de V11			Frecuencia de V14		
[0, 20609.59]	**FREC**	**PROBAB**	**[0, 6871.68]**	**FREC**	**PROBAB**
0	4823.00	0.33	0	13876.00	0.95
1717.47	5785.00	0.40	572.64	546.00	0.04
3434.93	1449.00	0.10	1145.28	70.00	0.00
5152.40	797.00	0.05	1717.92	23.00	0.00
6869.86	327.00	0.02	2290.56	11.00	0.00
8587.33	355.00	0.02	2863.20	5.00	0.00
10304.80	271.00	0.02	3435.84	2.00	0.00
12022.26	136.00	0.01	4008.48	1.00	0.00
13739.73	168.00	0.01	4581.12	0	0
15457.19	155.00	0.01	5153.76	1.00	0.00
17174.66	60.00	0.00	5726.40	0	0
18892.12	119.00	0.01	6299.04	0	0
20609.59	91.00	0.01	6871.68	1.00	0.00

Fuente: Elaboración propia

f) Rentabilidad en el clúster C2

- Total aceptados: 14,257 créditos
- Monto de crédito promedio: S/ 3,435.36
- Total monto de créditos otorgados: 48’977,928
- Plazo de pago promedio: 14 meses

- Intereses ganados: 9’673,697 soles
- Rentabilidad en el período de pago: 19.75%
- Rentabilidad anual: 16.9%

g) Riesgo en el clúster C2

- Máximo días de atraso de la última cuota pagada: 162 días
- Promedio días de atraso de la última cuota pagada: 2.84 días
- Riesgo: 1.75%

Conclusiones de la prueba

Las variables V2, V3, V11 y V14 tienen las más altas correlación en la formación de los clústeres, y las variables V8, V5 y V6 tienen la más alta correlación dentro de cada clúster.

(i) Los clientes agrupados en el clúster C1, presentan un monto del préstamo (V2) que se encuentra en el intervalo [20880, 297770] y se dispersan alrededor de la media de 40,272 soles, que se puede considerar como montos altos. En el clúster C2 están agrupados los clientes con monto de préstamo bajo, pues estos montos se encuentran en el intervalo [300, 35000] y se dispersan alrededor de la media de 3,435 soles. La intersección de los dos intervalos nos permite afirmar que en el intervalo [20880, 35000] hay clientes que pueden estar en el clúster C1 o en el clúster C2.

(ii) En el clúster C1 se agrupan los clientes cuyo saldo capital de la deuda (V3) tienen valores altos, pues estos se encuentran en el intervalo [17235, 297770] y se dispersan alrededor de la media de 38,058 soles. En el

clúster C2 están los clientes con saldo capital de la deuda bajo, pues estos se encuentran en el intervalo [16, 29822] y se dispersan alrededor de la media de 2,735 soles. Con la intersección de los dos intervalos se forma el intervalo [17235, 29822] en el que hay clientes que pueden estar en el clúster C1 o en el clúster C2.

(iii) Con respecto a la variable saldo de capital vigente de la operación (V11), en el clúster C1 están agrupados los clientes con saldos de capital que están en el intervalo [17235, 297770] con una media de 38,058 soles, considerado alto. En el clúster C2 se agrupan los clientes con saldos de capital que están en el intervalo [0, 20610] con una media de 2,688 soles, considerado como saldos de capital de monto bajo. De la intersección de los dos intervalos se forma el intervalo [17235, 20610] en el que hay clientes que pueden estar en el clúster C1 o el clúster C2.

(iv) Con respecto a la variable rendimientos devengados de la operación (V14), se aprecia que en el clúster C1 está agrupados los clientes cuyos valores de esta variable están en el intervalo [0, 19653] con una media de 1,456 soles, considerados como rendimiento devengado alto. En el clúster C2 están agrupados los clientes cuyo valor de esta variable se encuentra en el intervalo [0, 6872] con una media de 73 soles, estos valores son bajos con respecto a los valores que tienen los clientes en el clúster C1.

(v) Con el clúster C1 están agrupados los clientes que han tenido un retraso en el pago de su cuota en el intervalo [0, 60] con un promedio de 3.55 días de retraso; y el clúster C2 agrupa a los clientes que han tenido un retraso en el pago de su cuota que está en un intervalo mayor [0, 162] con un promedio ligeramente menor de 2.84 días. La variable días de atraso de la última cuota pagada (V8) tiene una correlación baja en la formación de los clústeres, pero dentro de los clústeres C1 y C2, es la

variable que tiene la más alta correlación (-0.78) con la variable de aceptación o rechazo de la solicitud de crédito.

(vi) El clúster C1 tiene una rentabilidad anual de 18.05% mayor que la del clúster C2 cuya rentabilidad es de 16.2%. El riesgo del clúster C1 es 5.92% también mayor 1.75% que es el riesgo del clúster C2.

(vii) La Tabla 27 resume estos resultados. Se concluye que, la red SOM ha agrupado en el clúster C1 los clientes con montos altos cuya rentabilidad y riesgo es mayor, y en el clúster C2 están agrupados los clientes con montos bajos, que a su vez a su vez tienen una rentabilidad y riesgo menor. No hay conjuntos disjuntos, porque la intersección de los conjuntos de los dos clústeres da lugar a un conjunto formado por clientes que pueden estar en cualquiera de los dos clústeres.

Tabla 27. Red SOM de dos neuronas con topología Gridtop y métrica Dist

Clúster	Tamaño	Aceptados	Rechazados	Valores de las variables	Rentabilidad	Riesgo
C1	1,033	999	34	Alto	18.05%	5.92%
C2	14,536	14,257	279	Bajo	16.9%	1.75%

Fuente: Elaboración propia

(II) Red Neuronal SOM: 2 neuronas, Hextop, Linkdist

En esta prueba se tienen dos clústeres C1 y C2 con 1,025 y 14,544 registros respectivamente, a continuación, se analizarán las características de los registros de cada clúster.

Variables influyentes en la formación de los clústeres

La Tabla 28 muestra las estadísticas de las 27 variables y su correlación con la formación de los clústeres.

Tabla 28. Estadísticas y CC con col 28 (de las 27 variables)

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.03	1	1	-0.1
V2	300	297770	10957.9	5879.45	1000	-0.84
V3	16.24	297770	10475.32	5078.65	1000	-0.84
V4	8	13	1.1	11.07	12	0.35
V5	0	4	0.59	0.15	0	0.03
V6	0	4	0.55	0.13	0	0.04
V7	-419	268	37.64	-14.98	-15	0.08
V8	0	162	8.32	2.88	0	-0.02
V9	-125	88	4.6	0.02	0	0
V10	0.83	48000	756.33	136.1	20	-0.24
V11	0	297770	10480.37	5034.86	1000	-0.84
V12	0	26527	453.83	38.05	0	0.02
V13	0	29822	350.02	5.75	0	0
V14	0	19652.81	498.09	164.71	0	-0.69
V15	0	5903	94.62	9.03	0	0
V16	40546	40847	84.95	40711.96	40751	0.06
V17	1	5	0.35	2.94	3	0.02
V18	0	360	21.54	6.35	0	-0.6
V19	40714	48131	438.13	41198.31	41001	-0.56
V20	40579	41209	37.64	40862.16	40849	-0.08
V21	1	365	35.29	35.84	30	-0.03
V22	1	240	14.3	15.65	12	-0.54
V23	0	37	2.77	3.47	0	0.07
V24	1	1	0	1	1	NaN
V25	1	44	13.38	18.98	1	-0.01
V26	12.01	293.79	18.25	51.09	58.27	0.39
V27	-1	1	0.28	0.96	1	0.02
Clústeres	1	2	0.25	1.93	2	1

Fuente: Elaboración propia

La Tabla 29 presenta un resumen de las variables que tienen la correlación más alta.

Tabla 29. Variables con la correlación más alta

Variable	Coeficiente de correlación	Descripción
V2	-0.84	Monto del crédito otorgado.
V3	-0.84	Saldo capital de la deuda.
V11	-0.84	Saldo de capital vigente de la operación.
V14	-0.69	Rendimientos devengados de la operación.

Fuente: Elaboración propia

Con respecto a las estadísticas y a la correlación se comenta lo siguiente:

(i) Se observa, por el valor de las correlaciones, las variables que más han influido en la formación de los clústeres son el monto del crédito otorgado (V2), el saldo capital de la deuda (V3), el saldo capital vigente de la operación (V11), y los rendimientos devengados de la operación (V14). La alta correlación que tienen estas variables explica porque son las que han determinado la formación de los clústeres.

(ii) Estas variables presentan valores grandes comparado con las otras variables, al observar la media de las otras variables, los valores de estas variables son altas, en los rangos, los valores máximos son altos. Lo que marca la contribución de una variable en el proceso de agrupamiento es la media, y este valor es alto en estas variables.

(iii) La contribución de estas variables en la formación de los clústeres, se explica porque se están usando distancias de enlace entre los datos del cliente con el centro de las neuronas y la topología es de las mallas hexagonales, y los valores numéricamente grandes están influyendo en la formación de los clústeres.

(iv) En el campo de las aplicaciones, la red neuronal ha clasificado a los clientes en función del monto del préstamo.

(v) La variable de aceptación o rechazo de la solicitud de crédito (V27) ha tenido una participación muy baja en la formación de los clústeres, esto se explica por el valor de su coeficiente de correlación (0.02).

Estudio del clúster C1

a) Correlaciones en el clúster C1

La Tabla 30 muestra las medidas estadísticas de los registros del clúster C1 y la correlación con V27.

Tabla 30. Estadísticas del clúster C1 (CC con V27)

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.12	1.01	1	0.02
V2	21560	297770	17286.16	40411.56	40000	0.01
V3	17380.95	297770	17136.25	38196.6	41625.02	0.01
V4	8	13	1.08	9.62	9	0.09
V5	0	3	0.3	0.07	0	-0.44
V6	0	1	0.23	0.06	0	-0.56
V7	-419	30	39.31	-26.5	-15	-0.2
V8	0	60	8.82	3.57	0	-0.78
V9	-18	31	1.65	-0.03	0	-0.11
V10	254.46	48000	2250.22	813.31	600	-0.08
V11	17380.95	297770	17136.25	38196.6	41625.02	0.01
V12	0	0	0	0	0	NaN
V13	0	0	0	0	0	NaN
V14	0	19652.81	1221.81	1464.64	1204.62	-0.17
V15	0	2791	129.55	7.27	0	0.01
V16	40547	40847	86.78	40693.38	40786	0.16
V17	1	3	0.41	2.91	3	-0.04
V18	0	120	56.37	55.32	120	-0.09
V19	40852	48131	787.45	42128.57	42700	-0.03
V20	40817	41209	38.78	40873.56	40862	0.2

V21	30	365	45.59	39.72	30	0.04
V22	1	240	26.16	44.97	60	-0.05
V23	0	9	2.28	2.75	0	-0.04
V24	1	1	0	1	1	NaN
V25	1	43	10.91	19.29	15	0.02
V26	12.01	49.36	7.05	24.1	18.02	0.02
V27	-1	1	0.36	0.93	1	1
Clúster C1	1	1	0	1	1	NaN

Fuente: Elaboración propia

Observaciones sobre el clúster C1:

(i) El clúster ha seleccionado clientes cuyo monto del crédito otorgado (V2), mayormente se concentran en alrededor de la media de 40,412 soles. Los créditos otorgados se encuentran en el intervalo [21560, 297770], y son pocos los créditos iguales o cerca al mínimo o máximo. La desviación estándar nos permite afirmar que hay una franja alrededor de la media en donde se encuentran el monto de préstamo de la mayoría de los clientes, y que estos oscilan entre los valores 23,125 soles y 57,698 soles. Esta variable ha tenido una participación significativa en la formación de los clústeres, esto se explica por el valor de su alta correlación (-0.84), en cambio, su correlación con la variable V27 es muy baja, lo que indica que V2 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(ii) Con respecto al saldo capital de la deuda (V3), el clúster ha seleccionado clientes cuyo saldo capital se concentran en alrededor de la media de 38,197 soles. El valor de esta variable fluctúa en el intervalo [17381, 297770], y hay un reducido número de clientes que tienen un saldo capital de la deuda cercanos o iguales al mínimo o al máximo. La desviación estándar nos permite afirmar que hay una franja alrededor de la media en donde se encuentran el saldo capital de la mayoría de los clientes, y que estos oscilan entre los valores 21,060 soles y 55,333 soles. Esta variable

también ha sido tomada en cuenta para la formación de los clústeres, esto se explica por el valor de su alta correlación (–0.84). En cambio, su correlación con la variable V27 es muy baja, lo que indica que V3 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iii) La variable saldo de capital vigente de la operación (V11) varía en el rango de 17,381 a 297,770 soles, su media es 38,197 soles. Hay un reducido número de clientes con un saldo de capital vigente de la operación con valores iguales o cercanos a los extremos, esto se explica por el valor de la media. Teniendo en cuenta la desviación estándar y la media, se observa que la mayoría de los clientes tienen un saldo de capital vigente de la operación que fluctúa entre los valores de 21,060 y 55,333 soles. La variable V11 también tiene un coeficiente de correlación (-0.84), lo que explica porque se ha tomado en cuenta en la formación de los clústeres. También es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que V11 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iv) La variable rendimientos devengados de la operación (V14) tiene valores que fluctúan desde un mínimo de cero hasta 19,653 soles. Teniendo en cuenta que la media es de 1,465 soles, se puede observar que hay un reducido número de clientes con valores altos de esta variable, la misma afirmación se puede hacer para el número de clientes que tienen valor cero para esta variable. Los rendimientos devengados de la operación de la mayoría de los clientes fluctúan entre los valores de 243 y 2,686 soles. La variable V14 tiene un coeficiente de correlación significativamente alto (-0.69), lo que explica porque se ha tomado en cuenta en la formación de los clústeres. Es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que V14 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(v) La variable días de atraso de la última cuota pagada (V8) varía en el rango de 0 a 60 días. Su media de 3.57 indica que la mayoría de los clientes ha pagado con un retraso de hasta 30 días, muy pocos son los clientes que han pagado con un retraso igual o cercano a los 60 días. Esta variable tiene una correlación baja en la formación de los clústeres, pero dentro del clúster C1, es la variable que tiene la más alta correlación (-0.78) con la variable de aceptación o rechazo de la solicitud de crédito (V27).

b) Frecuencia de las variables Aceptación o rechazo de la solicitud de crédito (V27) y Días de atraso de la última cuota pagada (V8) en el clúster C1

Tabla 31. Frecuencia de las variables V27 y V8

Frecuencia de V27			Frecuencia de V8		
-1 v 1	FREC	PROBAB	[0, 60]	FREC	PROBAB
-1.00	34.00	0.03	0	788.00	0.77
1.00	991.00	0.97	5.00	96.00	0.09
			10.00	39.00	0.04
			15.00	34.00	0.03
			20.00	15.00	0.01
			25.00	11.00	0.01
			30.00	14.00	0.01
			35.00	10.00	0.01
			40.00	4.00	0.00
			45.00	7.00	0.01
			50.00	2.00	0.00
			55.00	4.00	0.00
			60.00	1.00	0.00

Fuente: Elaboración propia

c) Frecuencia de las variables Clasificación del deudor (V5) y Clasificación del deudor sin considerar alineamiento del sistema (V6) en el clúster C1

Tabla 32. Frecuencia de las variables V5 y V6

Frecuencia de V5			Frecuencia de V6		
[0, 3]	FREC	PROBAB	[0, 1]	FREC	PROBAB
0	963.00	0.94	0	966.00	0.94
1.00	56.00	0.05	1.00	59.00	0.06
2.00	2.00	0.00			
3.00	4.00	0.00			

Fuente: Elaboración propia

Comentarios sobre la frecuencia de las variables V5 y V6 en el clúster C1:

Las variables V5 y V6, después de la variable V8, dentro del clúster C1 son las variables con la correlación más alta con respecto a la variable V27.

(i) Se observa que en el clúster C1, considerando el factor de la SBS para la clasificación de los deudores, el 94% de los clientes ha sido clasificado como un deudor normal, el 5% de los clientes ha sido clasificados como un cliente con potencial pérdida, y el resto ha sido clasificado como deficiente o dudoso.

(ii) Con respecto a la clasificación propia de la entidad financiera sin considerar el factor de la SBS, en este clúster el 94% de los clientes ha sido clasificado como un deudor normal, y el 6% ha sido clasificado como un cliente con potencial pérdida.

d) Frecuencia de las variables Monto del crédito otorgado (V2) y Saldo capital de la deuda (V3) en el clúster C1

Tabla 33. Frecuencia de las variables V2 y V3

Frecuencia de V2			Frecuencia de V3		
[21560, 297770]	FREC	PROBAB	[17380.95, 297770]	FREC	PROBAB
21560.00	265.00	0.26	17380.95	236.00	0.23
44577.50	712.00	0.69	40746.70	736.00	0.72
67595.00	18.00	0.02	64112.46	25.00	0.02
90612.50	17.00	0.02	87478.21	15.00	0.01
113630.00	8.00	0.01	110843.97	9.00	0.01
136647.50	2.00	0.00	134209.72	1.00	0.00
159665.00	1.00	0.00	157575.48	1.00	0.00
182682.50	0	0	180941.23	0	0
205700.00	0	0	204306.98	0	0
228717.50	0	0	227672.74	0	0
251735.00	0	0	251038.49	0	0
274752.50	1.00	0.00	274404.25	1.00	0.00
297770.00	1.00	0.00	297770.00	1.00	0.00

Fuente: Elaboración propia

e) Frecuencia de las variables Saldo capital vigente de la operación (V11) y Rendimientos devengados de la operación (V14) en el clúster C1

Tabla 34. Frecuencia de las variables V11 y V14

Frecuencia de V11			Frecuencia de V14		
[17380.95, 297770]	FREC	PROBAB	[0, 19652.81]	FREC	PROBAB
17380.95	236.00	0.23	0	418.00	0.41
40746.70	736.00	0.72	1637.73	384.00	0.37
64112.46	25.00	0.02	3275.47	212.00	0.21
87478.21	15.00	0.01	4913.20	7.00	0.01
110843.97	9.00	0.01	6550.94	2.00	0.00
134209.72	1.00	0.00	8188.67	0	0
157575.48	1.00	0.00	9826.41	1.00	0.00
180941.23	0	0	11464.14	0	0
204306.98	0	0	13101.87	0	0

227672.74	0	0	14739.61	0	0
251038.49	0	0	16377.34	0	0
274404.25	1.00	0.00	18015.08	0	0
297770.00	1.00	0.00	19652.81	1.00	0.00

Fuente: Elaboración propia

f) Rentabilidad en el clúster C1

- Total aceptados: 991 créditos
- Monto de crédito promedio: S/ 40,411.56
- Total monto de créditos otorgados: 40’047,856
- Tasa de interés: 30% anual
- Plazo de pago promedio: 45 meses
- Intereses ganados: 27’113,919
- Rentabilidad en el período de pago: 67.70%
- Rentabilidad anual: 18.05%

g) Riesgo en el clúster C1

- Máximo días de atraso de la última cuota pagada: 60 días
- Promedio días de atraso de la última cuota pagada: 3.57 días
- Riesgo: 5.95 %

Estudio del clúster C2

a) Correlaciones en el clúster C2

La Tabla 35 muestra las medidas estadísticas de los registros del clúster C1 y la correlación con V27.

Tabla 35. Estadísticas del clúster C2 (CC con V27)

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.02	1	1	0
V2	300	35000	4186.85	3445.77	1000	0.02
V3	16.24	29821.89	3746.84	2744.64	1000	0.02
V4	9	13	1.02	11.17	12	0.02
V5	0	4	0.6	0.15	0	-0.44
V6	0	4	0.57	0.14	0	-0.46
V7	-360	268	37.39	-14.17	-2	-0.27
V8	**0**	**162**	**8.28**	**2.84**	**0**	**-0.78**
V9	-125	88	4.74	0.03	0	-0.22
V10	0.83	29821.89	470.34	88.38	20	-0.19
V11	0	21000	3732.7	2697.76	1000	0.06
V12	0	26527	469.44	40.73	0	-0.23
V13	0	29822	362.14	6.15	0	-0.04
V14	0	6871.68	181.58	73.09	0	0.01
V15	0	5903	91.67	9.16	0	-0.22
V16	40546	40847	84.67	40713.27	40847	0.14
V17	1	5	0.35	2.94	3	-0.03
V18	0	360	9.6	2.9	0	0.02
V19	40714	47893	310.7	41132.75	41001	0.05
V20	40579	41207	37.43	40861.36	40849	0.27
V21	1	365	34.43	35.57	30	0.02
V22	1	240	10.29	13.58	12	0.01
V23	0	37	2.8	3.52	0	-0.03
V24	1	1	0	1	1	NaN
V25	1	44	13.54	18.95	1	0.01
V26	12.01	293.79	17.27	52.99	58.27	-0.06
V27	-1	1	0.27	0.96	1	1
Clúster C2	2	2	0	2	2	NaN

Fuente: Elaboración propia

Observaciones sobre el clúster C2:

(i) El clúster C2 ha seleccionado clientes cuyo monto del crédito otorgado (V2) es relativamente bajo, que están en el intervalo [300, 35000], estos créditos se concentran alrededor de la media de 3,446 soles, por tanto, hay un reducido número de clientes que tienen un monto del crédito cercano o igual al mínimo o al máximo. Esta variable ha tenido una participación importante en la formación de los clústeres, esto se explica por el valor de su correlación alta (-0.84), en cambio, su correlación con la variable V27 es muy baja, lo que indica que V2 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(ii) La variable saldo capital de la deuda (V3), indica que, en el período de repago del préstamo, el saldo del capital varía en el rango de 16 hasta 29,822 soles, siendo el promedio de 2,745 soles. Hay un reducido número de clientes que tienen un saldo capital de la deuda igual a los valores máximos o mínimo, esto se explica por el valor de la media. La mayoría de los clientes tienen un saldo de capital cuyo monto se encuentra alrededor de la media. El alto grado de correlación de esta variable (-0.84) explica porque esta variable ha sido tomada en cuenta para la formación de los clústeres. En cambio, su correlación con la variable V27 es muy baja, lo que indica que V3 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iii) La variable saldo de capital vigente de la operación (V11) varía en el intervalo [0, 21000], su media es 2,698 soles. Es reducido el número de clientes con un saldo de capital vigente de la operación con valores iguales o cercanos al máximo o al mínimo, es decir, la mayoría de los clientes tienen un saldo de capital vigente de la operación alrededor de la media. La variable V11 tiene un coeficiente de correlación alto cuyo

valores -0.84 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. También es necesario precisar que la correlación de esta variable con la variable V27 es muy baja, lo que indica que V11 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iv) La variable rendimientos devengados de la operación (V14) tiene valores que están en el intervalo [0, 6872]. Teniendo en cuenta que la media es de 73 soles, se puede afirmar que hay un reducido número de clientes con valores iguales o cercanos al mínimo o máximo, por lo que, para la mayoría de los clientes, el valor de esta variable se concentra alrededor de la media. La variable V14 tiene un coeficiente de correlación alto con valor -0.69 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. Es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que V14 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(v) La variable días de atraso de la última cuota pagada (V8) en el clúster C2 varía en el rango de 0 a 162 días. Su media de 2.84 días es ligeramente menor a la media de esta variable en el clúster C1. Esto indica que la mayoría de los clientes ha tenido un retraso de 3 días para pagar su cuota, solo algunos clientes se han pasado de los 30 días. Esta variable tiene una correlación baja en la formación de los clústeres, sin embargo, es la variable que tiene la mayor correlación con la variable de aceptación o rechazo de la solicitud de crédito.

b) Frecuencia de las variables Aceptación o rechazo de la solicitud de crédito (V27) y Días de atraso de la última cuota pagada (V8) en el clúster C2

Tabla 36. Frecuencia de las variables V27 y V8

Frecuencia de V27			Frecuencia de V8		
-1 v 1	FREC	PROBAB	[0, 162]	FREC	PROBAB
			0	12981.00	0.89
-1.00	279.00	0.02	13.50	1041.00	0.07
1.00	14265.00	0.98	27.00	285.00	0.02
			40.50	130.00	0.01
			54.00	63.00	0.00
			67.50	20.00	0.00
			81.00	13.00	0.00
			94.50	3.00	0.00
			108.00	3.00	0.00
			121.50	2.00	0.00
			135.00	2.00	0.00
			148.50	0	0
			162.00	1.00	0.00

Fuente: Elaboración propia

c) Frecuencia de las variables Clasificación del deudor (V5) y Clasificación del deudor sin considerar alineamiento del sistema (V6) en el clúster C2

Tabla 37. Frecuencia de las variables V5 y V6

Frecuencia de V5			Frecuencia de V6		
[0, 4]	FREC	PROBAB	[0, 4]	FREC	PROBAB
0	13407.00	0.92	0	13479.00	0.93
1.00	573.00	0.04	1.00	597.00	0.04
2.00	211.00	0.01	2.00	171.00	0.01
3.00	215.00	0.01	3.00	159.00	0.01
4.00	138.00	0.01	4.00	138.00	0.01

Fuente: Elaboración propia

Comentarios sobre la frecuencia de las variables V5 y V6 en el clúster C2:

Las variables V5 y V6, son las que tienen mayor con V27 después de la variable V8 dentro del clúster C2.

(i) Se observa que en el clúster C2, considerando el factor de la SBS para la clasificación de los deudores, el 92% de los clientes ha sido clasificado como un deudor normal, el 4% de los clientes ha sido clasificados como un cliente con potencial pérdida, el resto ha sido clasificado como deficiente, dudoso o pérdida.

(ii) Con respecto a la clasificación propia de la entidad financiera sin considerar el factor de la SBS, en este clúster el 93% de los clientes ha sido clasificado como un deudor normal, el 4% ha sido clasificado como un cliente con potencial pérdida, el resto ha sido clasificado como deficiente, dudoso o pérdida.

d) Frecuencia de las variables Monto del crédito otorgado (V2) y Saldo capital de la deuda (V3) en el clúster C2

Tabla 38. Frecuencia de las variables V2 y V3

Frecuencia de V2			**Frecuencia de V3**		
[300, 35000]	FREC	PROBAB	[16.24, 29821.89]	FREC	PROBAB
300.00	6652.00	0.46	16.24	7030.00	0.48
3191.67	4697.00	0.32	2500.04	4628.00	0.32
6083.33	1464.00	0.10	4983.85	1224.00	0.08
8975.00	791.00	0.05	7467.65	508.00	0.03
11866.67	184.00	0.01	9951.46	410.00	0.03
14758.33	358.00	0.02	12435.26	200.00	0.01
17650.00	64.00	0.00	14919.06	255.00	0.02
20541.67	301.00	0.02	17402.87	103.00	0.01
23433.33	10.00	0.00	19886.67	183.00	0.01

26325.00	9.00	0.00	22370.48	0	0
29216.67	12.00	0.00	24854.28	1.00	0.00
32108.33	1.00	0.00	27338.09	1.00	0.00
35000.00	1.00	0.00	29821.89	1.00	0.00

Fuente: Elaboración propia

e) Frecuencia de las variables Saldo capital vigente de la operación (V11) y Rendimientos devengados de la operación (V14) en el clúster C2

Tabla 39. Frecuencia de las variables V11 y V14

Frecuencia de V11			Frecuencia de V14		
[0, 20609.59]	FREC	PROBAB	[0, 6871.68]	FREC	PROBAB
0	5137.00	0.35	0	13879.00	0.95
1750.00	5558.00	0.38	572.64	550.00	0.04
3500.00	1428.00	0.10	1145.28	71.00	0.00
5250.00	750.00	0.05	1717.92	23.00	0.00
7000.00	345.00	0.02	2290.56	11.00	0.00
8750.00	338.00	0.02	2863.20	5.00	0.00
10500.00	262.00	0.02	3435.84	2.00	0.00
12250.00	145.00	0.01	4008.48	1.00	0.00
14000.00	182.00	0.01	4581.12	0	0
15750.00	127.00	0.01	5153.76	1.00	0.00
17500.00	80.00	0.01	5726.40	0	0
19250.00	176.00	0.01	6299.04	0	0
21000.00	16.00	0.00	6871.68	1.00	0.00

Fuente: Elaboración propia

f) Rentabilidad en el clúster C2

- Total aceptados: 14,265 créditos
- Monto de crédito promedio: S/ 3,445.77
- Total monto de créditos otorgados: 49’153,909

- Tasa de interés: 30% anual
- Plazo de pago promedio: 14 meses
- Intereses ganados: 9’708,455 soles
- Rentabilidad en el período de pago: 19.75%
- Rentabilidad anual: 16.9%

g) Riesgo en el clúster C2

- Máximo días de atraso de la última cuota pagada: 162 días
- Promedio días de atraso de la última cuota pagada: 2.84 días
- Riesgo: 1.75%

Conclusiones de la prueba

Las variables V2, V3, V11 y V14 tienen las más altas correlación en la formación de los clústeres, y las variables V8, V5 y V6 tienen la más alta correlación dentro de cada clúster.

(i) Los clientes agrupados en el clúster C1, tienen un monto del préstamo (V2) que se encuentra en el intervalo [21560, 297770] y se dispersan alrededor de la media de 40,412 soles, que se puede considerar como montos altos. En el clúster C2 están agrupados los clientes con monto de préstamo bajo, pues estos montos se encuentran en el intervalo [300, 35000] y se dispersan alrededor de la media de 3,446 soles. La intersección de los dos intervalos nos permite afirmar que en el intervalo [21560, 35000] hay clientes que pueden estar en el clúster C1 o en el clúster C2.

(ii) En el clúster C1 se agrupan los clientes cuyo saldo capital de la deuda (V3) tienen valores altos, pues estos se encuentran en el intervalo [17381, 297770] y se dispersan alrededor de la media de 38,197 soles. En el clúster C2 están los clientes con saldo capital de la deuda bajo, pues estos se encuentran en el intervalo [16, 29822] y se dispersan alrededor de la media de 2,745 soles. Con la intersección de los dos intervalos se forma el intervalo [17381, 29822] en el que hay clientes que pueden estar en el clúster C1 o en el clúster C2.

(iii) Con respecto a la variable saldo de capital vigente de la operación (V11), en el clúster C1 están agrupados los clientes con saldos de capital que están en el intervalo [17381, 297770] con una media de 38,197 soles, considerado alto. En el clúster C2 se agrupan los clientes con saldos de capital que están en el intervalo [0, 21000] con una media de 2,698 soles, considerado como saldos de capital de monto bajo. De la intersección de los dos intervalos se forma el intervalo [17381, 21000] en el que hay clientes que pueden estar en el clúster C1 o el clúster C2.

(iv) Con respecto a la variable rendimientos devengados de la operación (V14), se aprecia que en el clúster C1 está agrupados los clientes cuyos valores de esta variable están en el intervalo [0, 19653] con una media de 1,465 soles, considerados como rendimiento devengado alto. En el clúster C2 están agrupados los clientes cuyo valor de esta variable se encuentra en el intervalo [0, 6872] con una media de 73 soles, estos valores son bajos con respecto a los valores que tienen los clientes en el clúster C1.

(v) En el clúster C1 están agrupados los clientes que han tenido un retraso en el pago de su cuota en el intervalo [0, 60] con un promedio de 3.55 días de retraso; y el clúster C2 agrupa a los clientes que han tenido un retraso en el pago de su cuota que está en un intervalo mayor [0, 162] con un promedio ligeramente menor de 2.84 días. La variable días de

atraso de la última cuota pagada (V8) tiene una correlación baja en la formación de los clústeres, pero dentro de los clústeres C1 y C2, es la variable que tiene la más alta correlación (-0.78) con la variable de aceptación o rechazo de la solicitud de crédito.

(vi) El clúster C1 tiene una rentabilidad anual de 18.05% mayor que la del clúster C2 cuya rentabilidad es de 16.9%. El riesgo del clúster C1 es 5.95% también mayor 1.75% que es el riesgo del clúster C2.

(vii) Se resume los resultados en la Tabla 40 y se concluye que, la red SOM ha agrupado en el clúster C1 los clientes con montos altos cuya rentabilidad y riesgo es mayor, y en el clúster C2 están agrupados los clientes con montos bajos, que a su vez a su vez tienen una rentabilidad y riesgo menor. No hay conjuntos disjuntos, porque la intersección de los conjuntos de los dos clústeres da lugar a un conjunto formado por clientes que pueden estar en cualquiera de los dos clústeres.

Tabla 40. Red SOM de dos neuronas con topología Hextop y métrica Linkdist

Clúster	Tamaño	Aceptados	Rechazados	Valores de variables	Rentabilidad	Riesgo
C1	1,025	991	34	Alto	18.05%	5.95%
C2	14,544	14,265	279	Bajo	16.9%	1.75%

Fuente: Elaboración propia

(III) Red Neuronal SOM: 3 neuronas, Gridtop, Dist

En esta prueba se tienen tres clústeres C1, C2 y C3 con 13220, 1519 y 830 registros respectivamente, a continuación, se analizarán las características de los registros de cada clúster.

Variables influyentes en la formación de los clústeres

La Tabla 41 muestra las medidas estadísticas de las 27 variables y su coeficiente de correlación con la formación de los clústeres.

Tabla 41. Estadísticas y CC de la matriz de los 3 clústeres (C1, C2 y C3)

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.03	1	1	0.09
V2	300	297770	10957.9	5879.45	1000	0.88
V3	16.24	297770	10475.32	5078.65	1000	0.88
V4	8	13	1.1	11.07	12	-0.37
V5	0	4	0.59	0.15	0	-0.05
V6	0	4	0.55	0.13	0	-0.05
V7	-419	268	37.64	-14.98	-15	-0.14
V8	0	162	8.32	2.88	0	0
V9	-125	88	4.6	0.02	0	-0.01
V10	0.83	48000	756.33	136.1	20	0.26
V11	0	297770	10480.37	5034.86	1000	0.88
V12	0	26527	453.83	38.05	0	-0.01
V13	0	29822	350.02	5.75	0	0.02
V14	0	19652.81	498.09	164.71	0	0.70
V15	0	5903	94.62	9.03	0	0.01
V16	40546	40847	84.95	40711.96	40751	-0.03
V17	1	5	0.35	2.94	3	-0.09
V18	0	360	21.54	6.35	0	0.58
V19	40714	48131	438.13	41198.31	41001	0.65
V20	40579	41209	37.64	40862.16	40849	0.15
V21	1	365	35.29	35.84	30	0.11
V22	1	240	14.3	15.65	12	0.63
V23	0	37	2.77	3.47	0	-0.1
V24	1	1	0	1	1	NaN
V25	1	44	13.38	18.98	1	0.03
V26	12.01	293.79	18.25	51.09	58.27	-0.51
V27	-1	1	0.28	0.96	1	-0.01
Clústeres	1	3	0.52	1.2	1	1

Fuente: Elaboración propia

La Tabla 42 presenta las variables que tienen la correlación más alta en la formación de los 3 clústeres.

Tabla 42. Variables con la correlación más alta

Variable	Coeficiente de correlación	Descripción
V2	0.88	Monto del crédito otorgado.
V3	0.88	Saldo capital de la deuda.
V11	0.88	Saldo de capital vigente de la operación.
V14	0.70	Rendimientos devengados de la operación.

Fuente: Elaboración propia

Con respecto a las estadísticas y a la correlación mostrada en la Tabla 42 se comenta lo siguiente:

(i) Se observa, por el valor de las correlaciones, las variables que más han influido en la formación de los clústeres son el monto del crédito otorgado (V2), el saldo capital de la deuda (V3), el saldo capital vigente de la operación (V11) y los rendimientos devengados de la operación (V14). Estas cuatro variables tienen la más alta correlación y esto explica que estas variables son las que han determinado la formación de los clústeres. Es necesario indicar que el valor de las correlaciones de V2, V3 y V11 son superiores a los valores de correlación de estas mismas variables en las redes SOM con 2 clústeres.

(ii) Las variables V2, V3 y V11 presentan valores grandes comparado con las otras variables, la media es alta, en los rangos, los valores máximos son altos. Lo que marca la contribución de una variable en el proceso de agrupamiento es la media, y este valor es alto en estas variables.

(iii) La contribución de estas variables en la formación de los clústeres, se explica porque se están usando distancias euclidianas entre los datos del cliente con el centro de las neuronas y la topología es de las mallas

rectangulares, entonces los valores numéricamente grandes están influyendo en la formación de los clústeres.

(iv) En el campo de las aplicaciones, la red neuronal está privilegiando a los clientes que tienen préstamos más grandes, es decir ha clasificado en función del monto del préstamo.

(v) También es necesario comentar cómo ha sido la participación de la variable de aceptación o rechazo de la solicitud de crédito (V27) en la formación de los clústeres. El coeficiente de correlación de V27 de valor -0.01 indica que su participación ha sido muy baja.

Estudio del clúster C1

a) Correlaciones en el clúster C1

En la Tabla 43 se presenta las medidas estadísticas y la correlación de las variables con V27.

Tabla 43. Estadística del clúster C1 (CC con V27)

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.02	1	1	0
V2	300	20000	1906.5	2325.90	1000	0.01
V3	16.24	8581.21	1561.05	1724.38	1000	0.02
V4	9	13	0.99	11.22	12	0.02
V5	0	4	0.62	0.16	0	-0.44
V6	0	4	0.58	0.14	0	-0.46
V7	-360	268	34.04	-12.5	-15	-0.3
V8	**0**	**162**	**8.49**	**2.96**	**0**	**-0.78**
V9	-125	88	4.8	0.05	0	-0.22
V10	0.83	8581.21	234.52	63.25	20	-0.29
V11	0	8324.17	1572.03	1685.98	1000	0.08
V12	0	8581	295.33	37.94	0	-0.31
V13	0	6031	52.45	0.46	0	0
V14	0	1781.97	80.24	46.17	0	0.01

V15	0	1618	59.41	8.27	0	-0.3
V16	40546	40847	84.4	40712.41	40847	0.14
V17	1	5	0.3	2.96	3	-0.02
V18	0	117	6.74	2.48	0	0.03
V19	40714	45818	168.92	41087.08	41001	0.07
V20	40579	41207	34.08	40859.65	40849	0.3
V21	1	360	27.26	33.7	30	0.02
V22	1	168	5.57	12.19	12	0
V23	0	37	2.78	3.58	0	-0.02
V24	1	1	0	1	1	NaN
V25	1	44	13.52	18.7	1	0.01
V26	12.01	293.79	16.68	54.99	58.27	-0.05
V27	-1	1	0.28	0.96	1	1
Clúster C1	1	1	0	1	1	NaN

Fuente: Elaboración propia

Observaciones sobre el clúster C1:

(i) El clúster ha seleccionado clientes cuyo monto del crédito otorgado (V2), mayormente se concentran en alrededor de la media de 2,326 soles. El máximo crédito es de 20,000 soles, hay un reducido número de clientes que tienen ese valor máximo, esto se explica por el valor de la media. El monto mínimo del crédito en este clúster es de 300 soles, que es el otro extremo, y también lo tienen un reducido número de clientes. La desviación estándar nos permite afirmar que hay una franja alrededor de la media en donde se encuentran el monto del crédito de la mayoría de los clientes, y que estos oscilan entre los valores 419 soles y 4,232 soles. Por tanto, se puede afirmar que en el clúster C1 están los créditos con monto original relativamente bajo. Esta variable ha tenido una participación significativa en la formación de los clústeres, tiene un coeficiente de correlación alto (0.88), en cambio, su correlación con la variable V27 es muy baja, lo que indica que V2 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(ii) Con respecto al saldo capital de la deuda (V3), el clúster ha seleccionado clientes cuyo saldo capital se concentran en alrededor de la media de 1,724 soles. El monto máximo de saldo capital es de 8,581 soles, que lo tienen un reducido número de clientes, esto se explica por el valor de la media. El monto mínimo de saldo capital es de 16 soles, y también lo tienen un reducido número de clientes. La desviación estándar nos permite afirmar que hay una franja alrededor de la media en donde se encuentran el saldo capital de la mayoría de los clientes, y que estos oscilan entre los valores 163 soles y 3,285 soles. En este clúster están los clientes con saldo capital de la deuda relativamente bajo. Esta variable también ha sido tomada en cuenta para la formación de los clústeres, esto se explica por el valor de su alta correlación (0.88). En cambio, su correlación con la variable V27 es muy baja, lo que indica que V3 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iii) La variable saldo de capital vigente de la operación (V11) varía en el rango de 0 a 8,324 soles, su media es 1,686 soles. Hay un reducido número de clientes con un saldo de capital vigente de la operación con valores iguales o cercanos al máximo o al mínimo, esto se explica por el valor de la media. Teniendo en cuenta la desviación estándar y la media, se puede afirmar que la mayoría de los clientes tienen un saldo de capital vigente de la operación que fluctúa entre los valores de 114 y 3,258 soles, lo que se puede considerar un saldo capital vigente de la operación relativamente bajo. La variable V11 tiene un coeficiente de correlación más alto con valor 0.88 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. También es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que V11 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iv) La variable rendimientos devengados de la operación (V14) tiene valores que fluctúan desde un mínimo de cero hasta 1,782 soles. Teniendo en cuenta que la media es de 46 soles, se puede afirmar que hay un reducido número de clientes con valores altos de esta variable, también se puede afirmar que hay un reducido número de clientes con valor cero para esta variable. La variable V14 tiene un coeficiente de correlación significativamente alto con valor 0.70 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. Es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que V14 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(v) La variable días de atraso de la última cuota pagada (V8) varía en el rango de 0 a 162 días. Su media de 2.96 días indica que la mayoría de los clientes ha pagado hasta con 30 días de retraso, muy pocos son los clientes que han pagado con un retraso igual o cercano a 162 días. Esta variable tiene una correlación baja en la formación de los clústeres, pero es la que tiene la más alta correlación con la variable de aceptación o rechazo de la solicitud de crédito (V27).

b) Frecuencia de las variables Aceptación o rechazo de la solicitud de crédito (V27) y Días de atraso de la última cuota pagada (V8) en el clúster C1

Tabla 44. Frecuencia de las variables V27 y V8

Frecuencia de V27			Frecuencia de V8		
-1 ⩒ 1	FREC	PROBAB	[0 , 162]	FREC	PROBAB
-1.00	267.00	0.02	0	11738.00	0.89
1.00	12953.00	0.98	13.50	979.00	0.07
			27.00	278.00	0.02
			40.50	124.00	0.01

	54.00	60.00	0.00
	67.50	18.00	0.00
	81.00	12.00	0.00
	94.50	3.00	0.00
	108.00	3.00	0.00
	121.50	2.00	0.00
	135.00	2.00	0.00
	148.50	0	0
	162.00	1.00	0.00

Fuente: Elaboración propia

c) Frecuencia de las variables Clasificación del deudor (V5) y Clasificación del deudor sin considerar alineamiento del sistema (V6) en el clúster C1

Tabla 45. Frecuencia de las variables V5 y V6

Frecuencia de V5			Frecuencia de V6		
[0, 4]	FREC	PROBAB	[0, 4]	FREC	PROBAB
0	12129.00	0.92	0	12195.00	0.92
1.00	542.00	0.04	1.00	566.00	0.04
2.00	205.00	0.02	2.00	168.00	0.01
3.00	210.00	0.02	3.00	157.00	0.01
4.00	134.00	0.01	4.00	134.00	0.01

Fuente: Elaboración propia

Comentarios sobre la frecuencia de las variables V5 y V6 en el clúster C1:

(i) Se observa que en el clúster C1, considerando el factor de la SBS para la clasificación de los deudores, el 92% de los clientes ha sido clasificado como un deudor normal, el 4% de los clientes ha sido clasificados como un cliente con potencial pérdida, el 2% ha sido clasificado como deficiente, el 2% como dudoso, y el 1% como pérdida.

(ii) Con respecto a la clasificación propia de la entidad financiera sin considerar el factor de la SBS, en este clúster el 92% de los clientes ha sido clasificado como un deudor normal, el 4% ha sido clasificado como un cliente con potencial pérdida, y el resto ha sido clasificado como deficiente, dudoso o pérdida.

d) Frecuencia de las variables Monto del crédito otorgado (V2) y Saldo capital de la deuda (V3) en el clúster C1

Tabla 46. Frecuencia de las variables V2 y V3

Frecuencia de V2			Frecuencia de V3		
[300, 20000]	FREC	PROB	[16.24, 8581.21]	FREC	PROBAB
300.00	4684.00	0.35	16.24	869.00	0.07
1941.67	4778.00	0.36	729.99	5506.00	0.42
3583.33	1835.00	0.14	1443.73	2663.00	0.20
5225.00	1292.00	0.10	2157.48	1437.00	0.11
6866.67	224.00	0.02	2871.23	860.00	0.07
8508.33	251.00	0.02	3584.98	442.00	0.03
10150.00	139.00	0.01	4298.72	460.00	0.03
11791.67	8.00	0.00	5012.47	416.00	0.03
13433.33	0	0	5726.22	192.00	0.01
15075.00	6.00	0.00	6439.97	124.00	0.01
16716.67	0	0	7153.71	148.00	0.01
18358.33	1.00	0.00	7867.46	100.00	0.01
20000.00	2.00	0.00	8581.21	3.00	0.00

Fuente: Elaboración propia

e) Frecuencia de las variables Saldo capital vigente de la operación (V11) y Rendimientos devengados de la operación (V14) en el clúster C1

Tabla 47. Frecuencia de las variables V11 y V14

Frecuencia de V11			Frecuencia de V14		
[0, 8324.17]	FREC	PROBAB	[0, 1781.97]	FREC	PROBAB
0	1071.00	0.08	0	11079.00	0.84
693.68	5297.00	0.40	148.50	1852.00	0.14
1387.36	2625.00	0.20	297.00	181.00	0.01
2081.04	1442.00	0.11	445.49	41.00	0.00
2774.72	870.00	0.07	593.99	29.00	0.00
3468.40	393.00	0.03	742.49	13.00	0.00
4162.09	453.00	0.03	890.99	9.00	0.00
4855.77	476.00	0.04	1039.48	4.00	0.00
5549.45	151.00	0.01	1187.98	5.00	0.00
6243.13	162.00	0.01	1336.48	2.00	0.00
6936.81	145.00	0.01	1484.97	1.00	0.00
7630.49	84.00	0.01	1633.47	2.00	0.00
8324.17	51.00	0.00	1781.97	2.00	0.00

Fuente: Elaboración propia

f) Rentabilidad en el clúster C1

- Total créditos aceptados: 12,953
- Monto de crédito promedio: S/ 2,325.90
- Total monto de créditos otorgados: 30´127,383
- Plazo de pago promedio: 12 meses
- Intereses ganados: 4’704,858
- Rentabilidad en el período de pago: 15.6%
- Rentabilidad anual: 15.6%

g) Riesgo en el clúster C1

- Máximo días de atraso de la última cuota pagada: 162 días
- Promedio días de atraso de la última cuota pagada: 2.96 días
- Riesgo: 1.83 %

Estudio del clúster C2

a) Correlaciones en el clúster C2

La Tabla 48 muestra las medidas estadísticas de las variables y las correlaciones con la variable V27 en el clúster C2.

Tabla 48. Estadísticas del clúster C2 (CC con V27)

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.04	1	1	0
V2	8500	40000	5933.53	16227.70	10000	-0.08
V3	2928.69	29821.89	5245.08	14319.03	10000	-0.08
V4	9	13	1.23	10.63	10	0.03
V5	0	4	0.33	0.05	0	-0.43
V6	0	4	0.29	0.04	0	-0.49
V7	-355	210	56.28	-29.07	-2	-0.12
V8	**0**	**77**	**5.91**	**1.77**	**0**	**-0.79**
V9	-93	37	4.03	-0.15	0	-0.16
V10	43.93	29821.89	1258.32	348.54	150	-0.3
V11	0	28000	5307.18	14204.37	10000	0.02
V12	0	26527	1162.54	59.73	0	-0.28
V13	0	29822	1108.97	54.93	0	-0.17
V14	0	6871.68	484.22	368.6	0	-0.03
V15	0	5903	223	15.73	0	-0.24
V16	40546	40847	86.47	40720.03	40816	0.1
V17	1	3	0.61	2.8	3	-0.03
V18	0	360	22.08	7.01	0	0.02
V19	40797	48131	797.36	41609.4	41792	0.01
V20	40637	41202	56.24	40876.62	40849	0.12
V21	30	365	69.32	52.3	30	0.03
V22	1	240	27.09	28.27	24	-0.01

V23	0	11	2.85	3.03	0	-0.06
V24	1	1	0	1	1	NaN
V25	1	44	13.47	21.38	19	0.03
V26	12.01	54.65	7.59	32.69	29.84	-0.03
V27	-1	1	0.2	0.98	1	1
Clúster C2	2	2	0	2	2	NaN

Fuente: Elaboración propia

Observaciones sobre el clúster C2:

(i) El clúster C2 ha seleccionado clientes cuyo monto del crédito (V2) es mayor que en el clúster C1. Estos créditos se concentran alrededor de la media de 16,228 soles. El máximo crédito es de 40,000 soles, que lo tienen un reducido número de clientes, esto se explica por el valor de la media. El monto mínimo de crédito es de 8,500 soles, y también lo tienen un reducido número de clientes. Esta variable ha tenido una participación importante en la formación de los clústeres, esto se explica por el valor de su correlación alta (0.88), en cambio, su correlación con la variable V27 es muy baja, lo que indica que V2 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(ii) La variable saldo capital de la deuda (V3), indica que, en el período de repago del préstamo, el saldo del capital varía en el rango de 2,929 hasta 29,822 soles, siendo el promedio de 14,319 soles. Hay un reducido número de clientes que tienen un saldo capital de la deuda igual a los valores máximos o mínimo, esto se explica por el valor de la media. El alto grado de correlación de esta variable (0.88) explica porque esta ha sido tomada en cuenta para la formación de los clústeres. En cambio, su correlación con la variable V27 es muy baja, lo que indica que V3 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iii) La variable saldo de capital vigente de la operación (V11) varía en el rango de 0 a 28,000 soles, su media es 14,204 soles. Es reducido el número de clientes con un saldo de capital vigente de la operación con valores iguales o cercanos al máximo o al mínimo, es decir, la mayoría de los clientes tienen un saldo de capital vigente de la operación alrededor de la media, en el intervalo de 8,897 a 19,511 soles. La variable V11 tiene un coeficiente de correlación alto con valor 0.88 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. También es necesario precisar que la correlación de esta variable con la variable V27 es muy baja, lo que indica que V11 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iv) La variable rendimientos devengados de la operación (V14) tiene valores que fluctúan desde un mínimo de 0 hasta 6,872 soles. Teniendo en cuenta que la media es de 369 soles, se puede afirmar que hay un reducido número de clientes con valores iguales o cercanos al mínimo o máximo, por lo que, para la mayoría de los clientes, el valor de esta variable se concentra alrededor de la media. La variable V14 tiene un coeficiente de correlación significativamente alto con valor 0.70 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. Es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que V14 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(v) La variable días de atraso de la última cuota pagada (V8) en el clúster C2 varía en el rango de 0 a 77 días. Su media de 1.77 días es ligeramente menor a la media de esta variable en el clúster C1. Esto indica que la mayoría de los clientes ha tenido un retraso de 2 días para pagar su cuota, solo algunos clientes se han pasado de los 30 días. Esta variable tiene una correlación baja en la formación de los clústeres, sin embargo, es la

variable que tiene la mayor correlación con la variable de aceptación o rechazo de la solicitud de crédito.

b) Frecuencia de las variables Aceptación o rechazo de la solicitud de crédito (V27) y Días de atraso de la última cuota pagada (V8) en el clúster C2

Tabla 49. Frecuencia de las variables V27 y V8

Frecuencia de V27			Frecuencia de V8		
-1 v 1	FREC	PROBAB	[0, 77]	FREC	PROBAB
			0	1331.00	0.88
-1.00	15.00	0.01	6.42	122.00	0.08
1.00	504.00	0.99	12.83	34.00	0.02
			19.25	9.00	0.01
			25.67	6.00	0.00
			32.08	4.00	0.00
			38.50	4.00	0.00
			44.92	2.00	0.00
			51.33	4.00	0.00
			57.75	0	0
			64.17	1.00	0.00
			70.58	0	0
			77.00	2.00	0.00

Fuente: Elaboración propia

c) Frecuencia de las variables Clasificación del deudor (V5) y Clasificación del deudor sin considerar alineamiento del sistema (V6) en el clúster C2

Tabla 50. Frecuencia de las variables V5 y V6

Frecuencia de V5			Frecuencia de V6		
[0, 4]	FREC	PROBAB	[0, 4]	FREC	PROBAB
0	1468.00	0.97	0	1474.00	0.97
1.00	36.00	0.02	1.00	36.00	0.02
2.00	6.00	0.00	2.00	3.00	0.00
3.00	5.00	0.00	3.00	2.00	0.00
4.00	4.00	0.00	4.00	4.00	0.00

Fuente: Elaboración propia

Comentarios sobre la frecuencia de las variables V5 y V6 en el clúster C2:

Las variables V5 y V6, después de la variable V8, dentro del clúster C2 son las variables con la correlación más alta con respecto a la variable V27.

(i) Se observa que en el clúster C2, considerando el factor de la SBS para la clasificación de los deudores, el 97% de los clientes ha sido clasificado como un deudor normal, el 2% de los clientes ha sido clasificados como un cliente con potencial pérdida, y el resto ha sido clasificado como deficiente, dudoso o pérdida.

(ii) Con respecto a la clasificación propia de la entidad financiera sin considerar el factor de la SBS, en este clúster el 97% de los clientes ha sido clasificado como un deudor normal, el 2% ha sido clasificado como un cliente con potencial pérdida, y el resto ha sido clasificado como deficiente, dudoso o pérdida.

d) Frecuencia de las variables Monto del crédito otorgado (V2) y Saldo capital de la deuda (V3) en el clúster C2

Tabla 51. Frecuencia de las variables V2 y V3

Frecuencia de V2			Frecuencia de V3		
[8500, 40000]	FREC	PROBAB	[2928.69, 29821.89]	FREC	PROBAB
8500.00	22.00	0.01	2928.69	2.00	0.00
11125.00	514.00	0.34	5169.79	8.00	0.01
13750.00	358.00	0.24	7410.89	122.00	0.08
16375.00	52.00	0.03	9651.99	409.00	0.27
19000.00	323.00	0.21	11893.09	192.00	0.13
21625.00	38.00	0.03	14134.19	260.00	0.17
24250.00	107.00	0.07	16375.29	87.00	0.06
26875.00	20.00	0.01	18616.39	153.00	0.10
29500.00	71.00	0.05	20857.49	142.00	0.09
32125.00	2.00	0.00	23098.59	48.00	0.03
34750.00	10.00	0.01	25339.69	64.00	0.04
37375.00	0	0	27580.79	31.00	0.02
40000.00	2.00	0.00	29821.89	1.00	0.00

Fuente: Elaboración propia

e) Frecuencia de las variables Saldo capital vigente de la operación (V11) y Rendimientos devengados de la operación (V14) en el clúster C2

Tabla 52. Frecuencia de las variables V11 y V14

Frecuencia de V11			Frecuencia de V14		
[0, 28000]	FREC	PROBAB	[0, 6871.68]	FREC	PROBAB
0	8.00	0.01	0	880.00	0.58
2333.33	1.00	0.00	572.64	518.00	0.34
4666.67	8.00	0.01	1145.28	74.00	0.05
7000.00	58.00	0.04	1717.92	24.00	0.02
9333.33	465.00	0.31	2290.56	12.00	0.01
11666.67	182.00	0.12	2863.20	5.00	0.00
14000.00	273.00	0.18	3435.84	2.00	0.00

16333.33	92.00	0.06	4008.48	1.00	0.00
18666.67	152.00	0.10	4581.12	0	0
21000.00	142.00	0.09	5153.76	1.00	0.00
23333.33	51.00	0.03	5726.40	0	0
25666.67	63.00	0.04	6299.04	1.00	0.00
28000.00	24.00	0.02	6871.68	1.00	0.00

Fuente: Elaboración propia

f) Rentabilidad en el clúster C2

- Total aceptados: 1,504
- Monto de crédito promedio: S/ 16,227.70
- Total monto de créditos otorgados: 24’406,461
- Plazo de pago promedio: 28 meses
- Intereses ganados: 9’822,676
- Rentabilidad en el período de pago: 40.25 %
- Rentabilidad anual: 17.25 %

g) Riesgo en el clúster C2

- Máximo días de atraso de la última cuota pagada: 77 días
- Promedio días de atraso de la última cuota pagada: 1.77 días
- Riesgo: 2.3 %

Estudio del clúster C3

a) Correlaciones en el clúster C3

La Tabla 52 muestra las medidas estadísticas de las variables y su correlación con la variable V27 dentro del clúster C3.

Tabla 53. Estadísticas del clúster C3 (CC con V27)

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.12	1.02	1	0.02
V2	28500	297770	17743.34	43540.78	40000	0.03
V3	22860.75	297770	17339.07	41593.5	41625.02	0.03
V4	8	13	0.94	9.49	9	0.1
V5	0	3	0.33	0.08	0	-0.42
V6	0	1	0.25	0.07	0	-0.54
V7	-419	30	40.99	-28.63	-15	-0.21
V8	**0**	**60**	**9.19**	**3.77**	**0**	**-0.79**
V9	-18	19	1.18	-0.07	0	-0.01
V10	342.91	48000	2490.51	907.72	600	-0.07
V11	22860.75	297770	17339.07	41593.5	41625.02	0.03
V12	0	0	0	0	0	NaN
V13	0	0	0	0	0	NaN
V14	0	19652.81	1236.56	1679.59	1204.62	-0.17
V15	0	2791	143.93	8.98	0	0.01
V16	40547	40847	87.35	40690.17	40751	0.16
V17	1	3	0.41	2.91	3	-0.04
V18	0	120	56.64	66.74	120	-0.09
V19	40869	48084	662.14	42217.61	42700	-0.03
V20	40817	41209	40.38	40875.63	40862	0.21
V21	30	365	46.39	39.75	30	0.04
V22	1	240	22.31	47.62	60	-0.05
V23	0	9	2.13	2.58	0	-0.05
V24	1	1	0	1	1	NaN
V25	1	43	10.15	18.89	15	0.02
V26	12.55	49.36	6.35	22.62	18.02	0.01
V27	-1	1	0.38	0.93	1	1
Clúster C3	3	3	0	3	3	NaN

Fuente: Elaboración propia

Observaciones sobre el clúster C3:

(i) El clúster C3 ha seleccionado clientes cuyo monto del crédito otorgado (V2) es mayor que en el clúster C2 y mucho mayor que en el clúster C1. Estos créditos se concentran alrededor de la media de 43,541 soles. El monto máximo crédito es de 297,770 soles, que lo tienen un reducido número de clientes, esto se explica por el valor de la media. El monto mínimo de crédito es de 28,500 soles, y también lo tienen un reducido número de clientes. Esta variable ha tenido una participación importante en la formación de los clústeres, esto se explica por el valor de su correlación alta (0.88), en cambio, su correlación con la variable V27 es muy baja, lo que indica que V2 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(ii) La variable saldo capital de la deuda (V3), indica que, en el período de repago del préstamo, el saldo del capital varía en el rango de 22,861 hasta 297,770 soles, siendo el promedio de 41,594 soles. Hay un reducido número de clientes que tienen un saldo capital de la deuda igual a los valores máximos o mínimo, esto se explica por el valor de la media. La mayoría de los clientes tienen un saldo de capital cuyo monto se encuentra alrededor de la media, en el intervalo entre 24,255 y 58,933 soles. El alto grado de correlación de esta variable (0.88) explica porque esta ha sido tomada en cuenta para la formación de los clústeres. En cambio, su correlación con la variable V27 es muy baja, lo que indica que V3 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iii) La variable saldo de capital vigente de la operación (V11) varía en el rango de 22,861 a 297,770 soles, su media es 41,594 soles. Es reducido el número de clientes con un saldo de capital vigente de la operación con valores iguales o cercanos al máximo o al mínimo, es decir, la mayoría de los clientes tienen un saldo de capital vigente de la operación alrededor de la media. La variable V11 tiene un coeficiente de correlación alto con

valor 0.88 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. También es necesario precisar que la correlación de esta variable con la variable V27 es muy baja, lo que indica que esta variable no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iv) La variable rendimientos devengados de la operación (V14) tiene valores que fluctúan desde un mínimo de 0 hasta 19,653 soles. Teniendo en cuenta que la media es de 1,205 soles, se puede afirmar que hay un reducido número de clientes con valores iguales o cercanos al máximo, por lo que, para la mayoría de los clientes, el valor de esta variable se concentra alrededor de la media. Esta variable tiene un coeficiente de correlación significativamente alto con valor 0.70 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. Es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que esta variable no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(v) La variable días de atraso de la última cuota pagada (V8) en el clúster C3 varía en el rango de 0 a 60 días. Su media de 3.77 días es ligeramente mayor a la media de esta variable en el clúster C1. Esto indica que la mayoría de los clientes ha tenido un retraso de 4 días para pagar su cuota, solo algunos clientes se han pasado de los 30 días. Esta variable tiene una correlación baja en la formación de los clústeres, sin embargo, es la variable que tiene la mayor correlación con la variable de aceptación o rechazo de la solicitud de crédito.

b) Frecuencia de las variables Aceptación o rechazo de la solicitud de crédito (V27) y Días de atraso de la última cuota pagada (V8) en el clúster C3

Tabla 54. Frecuencia de las variables V27 y V8

Frecuencia de V27			Frecuencia de V8		
-1 v 1	FREC	PROBAB	[0, 60]	FREC	PROBAB
-1.00	31.00	0.04	0	636.00	0.77
1.00	799.00	0.96	5.00	74.00	0.09
			10.00	28.00	0.03
			15.00	31.00	0.04
			20.00	15.00	0.02
			25.00	9.00	0.01
			30.00	11.00	0.01
			35.00	10.00	0.01
			40.00	4.00	0.00
			45.00	6.00	0.01
			50.00	1.00	0.00
			55.00	4.00	0.00
			60.00	1.00	0.00

Fuente: Elaboración propia

c) Frecuencia de las variables Clasificación del deudor (V5) y Clasificación del deudor sin considerar alineamiento del sistema (V6) en el clúster C3

Tabla 55. Frecuencia de las variables V5 y V6

Frecuencia de V5			Frecuencia de V6		
[0, 3]	FREC	PROBAB	[0, 1]	FREC	PROBAB
0	773.00	0.93	0	776.00	0.93
1.00	51.00	0.06	1.00	54.00	0.07
2.00	2.00	0.00			
3.00	4.00	0.00			

Fuente: Elaboración propia

Comentarios sobre la frecuencia de las variables V5 y V6 en el clúster C3:

Las variables V5 y V6, después de la variable V8, dentro del clúster C3 son las variables con la correlación más alta con respecto a la variable V27.

(i) Se observa que en el clúster C3, considerando el factor de la SBS para la clasificación de los deudores, el 93% de los clientes ha sido clasificado como un deudor normal, el 3% de los clientes ha sido clasificados como un cliente con potencial pérdida, y el resto ha sido clasificado como deficiente o dudoso. Nadie ha sido clasificado como pérdida.

(ii) Con respecto a la clasificación propia de la entidad financiera sin considerar el factor de la SBS, en este clúster los clientes sólo han sido clasificados en dos categorías: el 93% de los clientes ha sido clasificado como un deudor normal y el 7% ha sido clasificado como un cliente con potencial pérdida. Nadie ha sido clasificado como deficiente, dudoso o pérdida.

d) Frecuencia de las variables Monto del crédito otorgado (V2) y Saldo capital de la deuda (V3) en el clúster C3

Tabla 56. Frecuencia de las variables V2 y V3

Frecuencia de V2			**Frecuencia de V3**		
[28500, 297770]	FREC	PROBAB	[22860.75, 297770]	FREC	PROBAB
28500.00	192.00	0.23	22860.75	174.00	0.21
50939.17	596.00	0.72	45769.85	611.00	0.74
73378.33	19.00	0.02	68678.96	23.00	0.03
95817.50	12.00	0.01	91588.06	12.00	0.01
118256.67	6.00	0.01	114497.17	6.00	0.01
140695.83	2.00	0.00	137406.27	1.00	0.00
163135.00	1.00	0.00	160315.38	1.00	0.00
185574.17	0	0	183224.48	0	0

208013.33	0	0	206133.58	0	0
230452.50	0	0	229042.69	0	0
252891.67	0	0	251951.79	0	0
275330.83	1.00	0.00	274860.90	1.00	0.00
297770.00	1.00	0.00	297770.00	1.00	0.00

Fuente: Elaboración propia

e) Frecuencia de las variables Saldo capital vigente de la operación (V11) y Rendimientos devengados de la operación (V14) en el clúster C3

Tabla 57. Frecuencia de las variables V11 y V14

Frecuencia de V11			**Frecuencia de V14**		
[22860.75, 297770]	FREC	PROBAB	[0, 19652.81]	FREC	PROBAB
22860.75	174.00	0.21	0	255.00	0.31
45769.85	611.00	0.74	1637.73	354.00	0.43
68678.96	23.00	0.03	3275.47	211.00	0.25
91588.06	12.00	0.01	4913.20	7.00	0.01
114497.17	6.00	0.01	6550.94	1.00	0.00
137406.27	1.00	0.00	8188.67	0	0
160315.38	1.00	0.00	9826.41	1.00	0.00
183224.48	0	0	11464.14	0	0
206133.58	0	0	13101.87	0	0
229042.69	0	0	14739.61	0	0
251951.79	0	0	16377.34	0	0
274860.90	1.00	0.00	18015.08	0	0
297770.00	1.00	0.00	19652.81	1.00	0.00

Fuente: Elaboración propia

f) Rentabilidad en el clúster C3

- Total aceptados: 799
- Monto de crédito promedio: S/ 43,540.78

- Total monto de créditos otorgados: 34’789,083
- Plazo de pago promedio: 48 meses
- Intereses ganados: 25’336,461
- Rentabilidad en el período de pago: 70.83 %
- Rentabilidad anual: 18.20 %

g) Riesgo en el clúster C3

- Máximo días de atraso de la última cuota pagada: 60 días
- Promedio días de atraso de la última cuota pagada: 3.77 días
- Riesgo: 6.28 %

Conclusiones de la prueba

Debido a que las variables V2, V3, V11 y V14 tienen la correlación más alta en la formación de los clústeres, luego de su estudio, se tienen las siguientes conclusiones:

(i) Los clientes agrupados en el clúster C1, tienen un monto del crédito otorgado (V2) que se encuentra en el intervalo [300, 20000], y tienen una media de 2,326 soles, se puede considerar como monto de crédito bajo. En el clúster C2 están los clientes con monto de préstamo moderadamente alto, pues estos se encuentran en el intervalo [8500, 40000] y se agrupan alrededor de la media de 16,228 soles. Es preciso indicar, que la intersección de estos dos intervalos da lugar a un conjunto donde hay clientes que pueden estar en el clúster C1 o en el clúster C2. En el clúster C3 están los clientes con monto del crédito en el intervalo [28500, 297770] y tienen una media de 43,541 soles, considerado monto del crédito alto. También se debe precisar que de la intersección de los

intervalos de los clústeres C2 y C3 se obtiene un conjunto donde hay clientes que pueden estar en el clúster C2 o en el clúster C3.

(ii) En el clúster C1 se agrupan los clientes cuyo saldo capital de la deuda (V3) están en el intervalo [16, 8581] y se agrupan alrededor de la media de 1,724 soles, que se puede considerar saldo de capital bajo. En el clúster C2 están los clientes con saldo capital de la deuda moderadamente alto, pues estos se encuentran en el intervalo [2929, 29822] y se agrupan alrededor de la media de 14,319 soles. Se debe precisar que la intersección de estos dos intervalos es un conjunto no vacío, en el que hay clientes pueden pertenecer al clúster C1 o al clúster C2. En el clúster C3 están agrupados los clientes con saldo capital de la deuda alto, pues se encuentran en el intervalo [22861, 297770] con una media de 41,594 soles. También es necesario precisar que de la intersección de los intervalos de los clústeres C2 y C3 se obtiene un conjunto donde hay clientes que pueden estar en el clúster C2 o en el clúster C3.

(iii) Con respecto a la variable saldo de capital vigente de la operación (V11), en el clúster C1 están agrupados los clientes con saldos de capital que se encuentran en el intervalo [0, 8324] y se dispersan alrededor de la media de 1,686 soles, considerado bajo. En el clúster C2 se agrupan los clientes con saldos de capital moderadamente alto, pues se encuentran en el intervalo [0, 28000] y se dispersan alrededor de la media de 14,204 soles. De la intersección de estos dos intervalos resulta un conjunto donde hay clientes que pueden estar en el clúster C1 o en el clúster C2. En el clúster C3 se agrupan los clientes con saldos de capital que están en el intervalo [22861, 297770] y cuya media es de 41,594 soles, considerado como saldos de capital vigente altos. También se indica que de la intersección de los intervalos de los clústeres C2 y C3 se obtiene un conjunto donde hay clientes que pueden estar en el clúster C2 o en el clúster C3.

(iv) Referente a la variable rendimientos devengados de la operación (V14), se aprecia que en el clúster C1 está agrupados los clientes con valores de esta variable en el intervalo [0, 1782] y cuya media es de 46 soles, considerado bajo. En el clúster C2 están agrupados los clientes cuyo valor de esta variable se encuentra en el intervalo [0, 6872] y se dispersan alrededor de la media de 369 soles, valor que es mayor que en el clúster C1. En el clúster C3 están agrupados los clientes cuyo valor de esta variable se encuentra en el intervalo [0, 19653] con una media de 1,680 soles, valor que es mayor que en el clúster C2.

(v) En el clúster C1, la variable días de atraso de la última cuota pagada (V8) varía en el rango de 0 a 162 días. Su media de 2.96 días indica que la mayoría de los clientes ha pagado con un retraso menor de 30 días, con un promedio de 3 días, son muy pocos son los clientes que han pagado con un retraso cercano o igual a 162 días. El clúster C2 agrupa a clientes que han tenido un retraso en el pago de su última cuota en el rango de 0 a 77 días, con una media de 1.77 días de retraso, ligeramente menor a la media de esta variable en el clúster C1. Por tanto, la mayoría de los clientes ha tenido un retraso de 2 días para pagar su cuota, solo algunos clientes se han pasado de 30 días de retraso. El clúster C3 agrupa a clientes cuyo atraso en el pago de su última cuota varía en el rango de 0 a 60 días, con una media de 3.77 días. Esto indica que la mayoría de los clientes ha tenido un retraso de 4 días para pagar su cuota, solo algunos clientes se han pasado de los 30 días. Esta variable (V8) tiene una correlación baja en la formación de los clústeres, pero tiene la más alta correlación con la variable de aceptación o rechazo de la solicitud de crédito (V27).

(vi) La rentabilidad anual en el clúster C1 es 15.6%, en el clúster C2 es 17.25%, y en el clúster C3 cuyo valor es 18.2%. El riesgo en el clúster C1 es 1.83%, en el clúster C2 es 2.3% y en el clúster C3 es 6.28%

(vii) Se concluye que, la red SOM ha clasificado en el clúster C1 al grupo más numeroso de clientes, cuyos valores de estas cuatro variables son bajos; la rentabilidad y el riesgo también son bajos en este clúster. En el clúster C2 están agrupados los clientes con valores de estas variables moderadamente altos; la rentabilidad y el riesgo en este clúster tienen valores mayores que el clúster C1. En el clúster C3 están agrupados los clientes con valores altos en estas cuatro variables, la rentabilidad es ligeramente mayor que en el clúster C2, pero el riesgo tiene un valor más alto.

(viii) No hay conjuntos disjuntos. Al intersectar dos clústeres contiguos se obtiene un conjunto que contiene clientes que pueden pertenecer a uno de los dos clústeres. En la Tabla 58 se resumen las características de los registros en cada clúster.

Tabla 58. Red SOM de tres neuronas con topología Gridtop y métrica Dist

Clúster	Tamaño	Aceptados	Rechazados	Valores de variables	Rentabilidad	Riesgo
C1	13,220	12,953	267	Bajo	15.60%	1.83%
C2	1,519	1,504	15	Medio	17.25%	2.30%
C3	830	799	31	Alto	18.20%	6.28%

Fuente: Elaboración propia

(IV) Red Neuronal SOM: 3 neuronas, Hextop, Linkdist

En esta prueba se tienen tres clústeres C1, C2 y C3 con 13218, 1528 y 823 registros respectivamente, a continuación, se analizarán las características de los registros de cada clúster.

Variables influyentes en la formación de los clústeres

En la Tabla 59 se muestran las medidas estadísticas de las todas las variables y el coeficiente de correlación que tienen con la formación de los clústeres.

Tabla 59. Estadísticas y CC de la matriz de los 3 clústeres (C1, C2 y C3)

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.03	1	1	0.09
V2	300	297770	10957.9	5879.45	1000	0.88
V3	16.24	297770	10475.32	5078.65	1000	0.88
V4	8	13	1.1	11.07	12	-0.37
V5	0	4	0.59	0.15	0	-0.05
V6	0	4	0.55	0.13	0	-0.05
V7	-419	268	37.64	-14.98	-15	-0.14
V8	0	162	8.32	2.88	0	0
V9	-125	88	4.6	0.02	0	-0.01
V10	0.83	48000	756.33	136.1	20	0.26
V11	0	297770	10480.37	5034.86	1000	0.88
V12	0	26527	453.83	38.05	0	-0.01
V13	0	29822	350.02	5.75	0	0.02
V14	0	19652.81	498.09	164.71	0	0.70
V15	0	5903	94.62	9.03	0	0.01
V16	40546	40847	84.95	40711.96	40751	-0.03
V17	1	5	0.35	2.94	3	-0.09
V18	0	360	21.54	6.35	0	0.58
V19	40714	48131	438.13	41198.31	41001	0.65
V20	40579	41209	37.64	40862.16	40849	0.15
V21	1	365	35.29	35.84	30	0.11
V22	1	240	14.3	15.65	12	0.63
V23	0	37	2.77	3.47	0	-0.1
V24	1	1	0	1	1	NaN
V25	1	44	13.38	18.98	1	0.03
V26	12.01	293.79	18.25	51.09	58.27	-0.51
V27	-1	1	0.28	0.96	1	-0.01
Clústeres	1	3	0.52	1.2	1	1

Fuente: Elaboración propia

La clusterización clasifica los clientes en 3 grupos, de acuerdo a sus datos. La Tabla 50 presenta un resumen de las variables que tienen la correlación más alta.

Tabla 60. Variables con coeficientes de correlación más alto

Variable	Coeficiente de correlación	Descripción
V2	0.88	Monto del crédito otorgado.
V3	0.88	Saldo capital de la deuda.
V11	0.88	Saldo de capital vigente de la operación.
V14	0.70	Rendimientos devengados de la operación.

Fuente: Elaboración propia

Observaciones:

(i) Se observa, por el valor de las correlaciones, las variables que más han influido en la formación de los clústeres son el monto del crédito otorgado (V2), el saldo capital de la deuda (V3), el saldo capital vigente de la operación (V11) y los rendimientos devengados de la operación (V14). Estas cuatro variables tienen la más alta correlación y esto explica que estas variables son las que han determinado la formación de los clústeres. Es necesario indicar que las medidas estadísticas y el valor de las correlaciones son iguales a los valores obtenidos para la red SOM con 3 neuronas con topología GridTop y métrica Dist.

(ii) Las variables V2, V3 y V11 presentan valores grandes comparado con las otras variables, la media es alta, en los rangos, los valores máximos son altos. Lo que marca la contribución de una variable en el proceso de agrupamiento es la media, y este valor es alto en estas variables.

(iii) La contribución de estas variables en la formación de los clústeres, se explica porque se están usando distancias de enlace entre los datos del cliente con el centro de las neuronas, y la topología es de las mallas

hexagonales, entonces los valores numéricamente grandes están influyendo en la formación de los clústeres.

(iv) En el campo de las aplicaciones, la red neuronal está privilegiando a los clientes que tienen préstamos más grandes, es decir ha clasificado en función del monto del préstamo.

(v) También es necesario comentar cómo ha sido la participación de la variable de aceptación o rechazo de la solicitud de crédito (V27) en la formación de los clústeres. El coeficiente de correlación de V27 de valor -0.01 indica que su participación ha sido muy baja.

Estudio del clúster C1

a) Correlaciones en el clúster C1

La Tabla 61 muestra las medidas estadísticas de las variables y su correlación con la variable V27 dentro del clúster C1.

Tabla 61. Estadísticas del clúster C1 (CC con V27)

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.02	1	1	0
V2	300	20000	1904.95	2324.91	1000	0.01
V3	16.24	8581.21	1559.1	1723.4	1000	0.02
V4	9	13	0.99	11.22	12	0.02
V5	0	4	0.62	0.16	0	-0.44
V6	0	4	0.58	0.14	0	-0.46
V7	-360	268	34.04	-12.5	-15	-0.3
V8	**0**	**162**	**8.49**	**2.96**	**0**	**-0.78**
V9	-125	88	4.8	0.05	0	-0.22
V10	0.83	8581.21	234.54	63.24	20	-0.29
V11	0	8293.71	1570.08	1684.99	1000	0.08
V12	0	8581	295.35	37.95	0	-0.31

V13	0	6031	52.46	0.46	0	0
V14	0	1781.97	80.23	46.15	0	0.01
V15	0	1618	59.41	8.27	0	-0.3
V16	40546	40847	84.41	40712.4	40847	0.14
V17	1	5	0.3	2.96	3	-0.02
V18	0	117	6.74	2.48	0	0.03
V19	40714	45818	168.77	41086.99	41001	0.07
V20	40579	41207	34.08	40859.65	40849	0.3
V21	1	360	27.26	33.7	30	0.02
V22	1	168	5.56	12.19	12	0
V23	0	37	2.78	3.58	0	-0.02
V24	1	1	0	1	1	NaN
V25	1	44	13.52	18.71	1	0.01
V26	12.01	293.79	16.68	55	58.27	-0.05
V27	-1	1	0.28	0.96	1	1
Clúster C1	1	1	0	1	1	NaN

Fuente: Elaboración propia

Observaciones sobre el clúster C1:

(i) El clúster ha seleccionado clientes cuyo monto de crédito (V2), mayormente se concentran en alrededor de la media de 2,325 soles. El monto de crédito varía en el intervalo [300, 20000], hay un reducido número de clientes con montos iguales o cercanos al máximo o al mínimo, esto se explica por el valor de la media. La desviación estándar nos permite afirmar que hay una franja alrededor de la media en donde se encuentran el monto de préstamo de la mayoría de los clientes, y que estos oscilan entre los valores 420 soles y 4,230 soles. Por tanto, se puede afirmar que en el clúster C1 están los créditos con monto relativamente bajo. Esta variable tiene un coeficiente de correlación alto (0.88), por lo tanto, ha tenido una participación significativa en la formación de los clústeres. En cambio, su correlación con la variable V27 es muy baja, lo que indica que V2 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(ii) Con respecto al saldo capital de la deuda (V3), los valores de esta variable se encuentran en el intervalo [16.24, 8581.21], y son pocos los clientes con saldo capital iguales o cercanos a los valores máximo o mínimo, esto se explica por la media cuyo valor es de 1,723 soles. La desviación estándar nos permite afirmar que hay una franja alrededor de la media en donde se encuentran el saldo capital de la mayoría de los clientes, y que estos oscilan entre los valores 163 soles y 3,285 soles, por lo tanto, en este clúster están los clientes con saldo capital de la deuda relativamente bajo. Esta variable también ha influido significativamente en la formación de los clústeres, esto se explica por el valor de su alta correlación (0.88). En cambio, su correlación con la variable V27 es muy baja, lo que indica que V3 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iii) La variable saldo de capital vigente de la operación (V11) tiene un mínimo de 0 soles y un máximo de 8,294 soles, su media es de 1,685 soles. Hay un reducido número de clientes con un saldo de capital vigente de la operación con valores iguales o cercanos a los extremos, esto se explica por el valor de la media. Teniendo en cuenta la desviación estándar y la media, se puede afirmar que la mayoría de los clientes tienen un saldo de capital vigente de la operación que fluctúa entre los valores de 114 y 3,256 soles, lo que se puede considerar un saldo capital vigente de la operación relativamente bajo. La variable V11 tiene un coeficiente de correlación más alto con valor 0.88 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. También es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que V11 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iv) La variable rendimientos devengados de la operación (V14) tiene valores que fluctúan desde un mínimo de 0 hasta 1,782 soles. Teniendo en cuenta

que la media es de 46 soles, se puede afirmar que hay un reducido número de clientes con valores cercanos o iguales al mínimo o al máximo. Esta variable tiene un coeficiente de correlación significativamente alto con valor 0.70 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. Es necesario precisar que su correlación con la variable V27 es muy baja, lo que indica que no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(v) La variable días de atraso de la última cuota pagada (V8) varía en el rango de 0 a 162 días. Su media de 2.96 días indica que la mayoría de los clientes ha pagado con un retraso de hasta 30 días, muy pocos son los clientes que han pagado con un retraso igual o cercano a días. Esta variable tiene una correlación baja en la formación de los clústeres, pero es la que tiene la más alta correlación con la variable de aceptación o rechazo de la solicitud de crédito (V27).

b) Frecuencia de las variables Aceptación o rechazo de la solicitud de crédito (V27) y Días de atraso de la última cuota pagada (V8) en el clúster C1

Tabla 62. Frecuencia de las variables V27 y V8

Frecuencia de V27			Frecuencia de V8		
-1 ⩒ 1	FREC	PROBAB	[0 , 162]	FREC	PROBAB
-1.00	267.00	0.02	0	11736.00	0.89
1.00	12951.00	0.98	13.50	979.00	0.07
			27.00	278.00	0.02
			40.50	124.00	0.01
			54.00	60.00	0.00
			67.50	18.00	0.00
			81.00	12.00	0.00
			94.50	3.00	0.00
			108.00	3.00	0.00

	121.50	2.00	0.00
	135.00	2.00	0.00
	148.50	0	0
	162.00	1.00	0.00

Fuente: Elaboración propia

c) Frecuencia de las variables Clasificación del deudor (V5) y Clasificación del deudor sin considerar alineamiento del sistema (V6) en el clúster C1

Tabla 63. Frecuencia de las variables V5 y V6

Frecuencia de V5			Frecuencia de V6		
[0, 4]	FREC	PROBAB	[0, 4]	FREC	PROBAB
0	12127.00	0.92	0	12193.00	0.92
1.00	542.00	0.04	1.00	566.00	0.04
2.00	205.00	0.02	2.00	168.00	0.01
3.00	210.00	0.02	3.00	157.00	0.01
4.00	134.00	0.01	4.00	134.00	0.01

Fuente: Elaboración propia

Comentarios sobre la frecuencia de las variables V5 y V6 en el clúster C1:

Las variables V5 y V6, después de la variable V8, dentro del clúster C1 son las variables con la correlación más alta con respecto a la variable V27.

(i) Se observa que en el clúster C1, considerando el factor de la SBS para la clasificación de los deudores, el 92% de los clientes ha sido clasificado como un deudor normal, el 4% de los clientes ha sido clasificados como un cliente con potencial pérdida, el 2% ha sido clasificado como deficiente, el 2% como dudoso, y el 1% como pérdida.

(ii) Con respecto a la clasificación propia de la entidad financiera sin considerar el factor de la SBS, en este clúster el 92% de los clientes ha sido clasificado como un deudor normal, el 4% ha sido clasificado como un cliente con potencial pérdida, y el resto ha sido clasificado como deficiente, dudoso o pérdida.

d) Frecuencia de las variables Monto del crédito otorgado (V2) y Saldo capital de la deuda (V3) en el clúster C1

Tabla 64. Frecuencia de las variables V2 y V3

Frecuencia de V2			Frecuencia de V3		
[300, 20000]	FREC	PROBAB	[16.24, 8581.21]	FREC	PROBAB
300.00	4684.00	0.35	16.24	869.00	0.07
1941.67	4778.00	0.36	729.99	5506.00	0.42
3583.33	1835.00	0.14	1443.73	2663.00	0.20
5225.00	1292.00	0.10	2157.48	1437.00	0.11
6866.67	224.00	0.02	2871.23	860.00	0.07
8508.33	249.00	0.02	3584.98	442.00	0.03
10150.00	139.00	0.01	4298.72	460.00	0.03
11791.67	8.00	0.00	5012.47	416.00	0.03
13433.33	0	0	5726.22	192.00	0.01
15075.00	6.00	0.00	6439.97	124.00	0.01
16716.67	0	0	7153.71	148.00	0.01
18358.33	1.00	0.00	7867.46	99.00	0.01
20000.00	2.00	0.00	8581.21	2.00	0.00

Fuente: Elaboración propia

e) Frecuencia de las variables Saldo capital vigente de la operación (V11) y Rendimientos devengados de la operación (V14) en el clúster C1

Tabla 65. Frecuencia de las variables V11 y V14

Frecuencia de V11			Frecuencia de V14		
[0, 8293.71]	FREC	PROBAB	[0, 1781.97]	FREC	PROBAB
0	1049.00	0.08	0	11078.00	0.84
691.14	5309.00	0.40	148.50	1851.00	0.14
1382.28	2575.00	0.19	297.00	181.00	0.01
2073.43	1499.00	0.11	445.49	41.00	0.00
2764.57	866.00	0.07	593.99	29.00	0.00
3455.71	392.00	0.03	742.49	13.00	0.00
4146.85	450.00	0.03	890.99	9.00	0.00
4838.00	483.00	0.04	1039.48	4.00	0.00
5529.14	151.00	0.01	1187.98	5.00	0.00
6220.28	163.00	0.01	1336.48	2.00	0.00
6911.42	146.00	0.01	1484.97	1.00	0.00
7602.57	86.00	0.01	1633.47	2.00	0.00
8293.71	49.00	0.00	1781.97	2.00	0.00

Fuente: Elaboración propia

f) Rentabilidad en el clúster C1

- Total aceptados: 12,951 créditos
- Monto de crédito promedio: S/ 2,324.91
- Total monto de créditos otorgados: S/ 30’109,909
- Tasa de interés: 30% anual
- Plazo de pago promedio: 12 meses
- Intereses ganados: 5’114,033
- Rentabilidad en el período de pago: 16.98 %
- Rentabilidad anual: 16.98 %

g) Riesgo en el clúster C1

- Máximo días de atraso de la última cuota pagada: 162 días
- Promedio días de atraso de la última cuota pagada: 2.96 días
- Riesgo: 1.83 %

Estudio del clúster C2

a) Correlaciones en el clúster C2

La Tabla 66 muestra las medidas estadísticas de las variables y su correlación con la variable V27 dentro del clúster C2.

Tabla 66. Estadísticas del clúster C2 (CC con V27)

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.04	1	1	0
V2	8500	40000	5991.2	16279.50	10000	-0.08
V3	2928.69	29821.89	5314.31	14373.37	10000	-0.07
V4	9	13	1.23	10.63	10	0.03
V5	0	4	0.33	0.05	0	-0.43
V6	0	4	0.29	0.04	0	-0.49
V7	-355	210	56.25	-29.05	-2	-0.12
V8	**0**	**77**	**5.9**	**1.76**	**0**	**-0.79**
V9	-93	37	4.02	-0.15	0	-0.16
V10	43.93	29821.89	1254.66	348.88	150	-0.3
V11	0	28500	5376.41	14259.38	10000	0.02
V12	0	26527	1159.12	59.38	0	-0.28
V13	0	29822	1105.71	54.61	0	-0.17
V14	0	6871.68	483.15	368.9	0	-0.03
V15	0	5903	222.35	15.63	0	-0.24
V16	40546	40847	86.35	40720.26	40816	0.1
V17	1	3	0.61	2.8	3	-0.03
V18	0	360	22.02	6.99	0	0.02
V19	40797	48131	795.81	41609.05	41792	0.01

V20	40637	41202	56.21	40876.6	40849	0.12
V21	30	365	69.21	52.27	30	0.03
V22	1	240	27.04	28.26	24	-0.01
V23	0	11	2.85	3.02	0	-0.06
V24	1	1	0	1	1	NaN
V25	1	44	13.46	21.35	19	0.03
V26	12.01	54.65	7.6	32.66	29.84	-0.03
V27	-1	1	0.2	0.98	1	1
Clúster C2	2	2	0	2	2	NaN

Fuente: Elaboración propia

Observaciones sobre el clúster C2:

(i) El clúster C2 ha seleccionado clientes cuyo monto del crédito (V2) es mayor que en el clúster C1. Estos créditos se concentran alrededor de la media de 16,280 soles. El monto del crédito varía en el intervalo [8500, 40000], y son pocos los clientes que tienen un crédito igual o cercano al mínimo o máximo, esto se explica por el valor de la media. Esta variable ha tenido una participación importante en la formación de los clústeres, esto se explica por el valor de su correlación alta (0.88), en cambio, su correlación con la variable V27 es muy baja, lo que indica que V2 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(ii) La variable saldo capital de la deuda (V3), indica que, en el período de repago del préstamo, el saldo del capital varía en el rango de 2,929 hasta 29,822 soles, siendo el promedio de 14,373 soles. Hay un reducido número de clientes que tienen un saldo capital de la deuda igual a los valores máximos o mínimo, la mayoría de los clientes tienen un saldo de capital cuyo monto se encuentra alrededor de la media. El alto grado de correlación de esta variable (0.88) explica porque ha sido tomada en cuenta para la formación de los clústeres. En cambio, su correlación con la variable V27 es muy baja, lo que indica que V3 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iii) La variable saldo de capital vigente de la operación (V11) varía en el rango de 0 a 28,500 soles, su media es 14,259 soles. Es reducido el número de clientes con un saldo de capital vigente de la operación con valores iguales o cercanos al máximo o al mínimo, es decir, la mayoría de los clientes tienen un saldo de capital vigente de la operación alrededor de la media, en el intervalo de 8,883 a 19,636 soles. La variable V11 tiene un coeficiente de correlación alto con valor 0.88 lo que explica su participación en la formación de los clústeres. También es necesario precisar que la correlación de esta variable con la variable V27 es muy baja, lo que indica que V11 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iv) La variable rendimientos devengados de la operación (V14) tiene valores que fluctúan desde un mínimo de 0 hasta 6,872 soles. Teniendo en cuenta que la media es de 369 soles, se puede afirmar que hay un reducido número de clientes con valores iguales o cercanos al mínimo o máximo, para la mayoría de los clientes, el valor de esta variable se concentra alrededor de la media. La variable V14 tiene un coeficiente de correlación alto con valor 0.70 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. Es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que V14 no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(v) La variable días de atraso de la última cuota pagada (V8) en el clúster C2 varía en el rango de 0 a 77 días. Su media de 1.76 días es ligeramente menor a la media de esta variable en el clúster C1. Esto indica que la mayoría de los clientes ha tenido un retraso de 2 días para pagar su cuota, solo algunos clientes se han pasado de los 30 días. Esta variable tiene una correlación baja en la formación de los clústeres, sin embargo, es la variable que tiene la mayor correlación con la variable de aceptación o rechazo de la solicitud de crédito.

b) Frecuencia de las variables Aceptación o rechazo de la solicitud de crédito (V27) y Días de atraso de la última cuota pagada (V8) en el clúster C2

Tabla 67. Frecuencia de las variables V27 y V8

Frecuencia de V27			Frecuencia de V8		
-1 v 1	FREC	PROBAB	[0, 77]	FREC	PROBAB
			0	1340.00	0.88
-1.00	15.00	0.01	6.42	122.00	0.08
1.00	1513.00	0.99	12.83	34.00	0.02
			19.25	9.00	0.01
			25.67	6.00	0.00
			32.08	4.00	0.00
			38.50	4.00	0.00
			44.92	2.00	0.00
			51.33	4.00	0.00
			57.75	0	0
			64.17	1.00	0.00
			70.58	0	0
			77.00	2.00	0.00

Fuente: Elaboración propia

c) Frecuencia de las variables Clasificación del deudor (V5) y Clasificación del deudor sin considerar alineamiento del sistema (V6) en el clúster C2

Tabla 68. Frecuencia de las variables V5 y V6

Frecuencia de V5			Frecuencia de V6		
[0, 4]	FREC	PROBAB	[0, 4]	FREC	PROBAB
0	1477.00	0.97	0	1483.00	0.97
1.00	36.00	0.02	1.00	36.00	0.02
2.00	6.00	0.00	2.00	3.00	0.00
3.00	5.00	0.00	3.00	2.00	0.00
4.00	4.00	0.00	4.00	4.00	0.00

Fuente: Elaboración propia

Comentarios sobre la frecuencia de las variables V5 y V6 en el clúster C2:

Las variables V5 y V6, después de la variable V8, dentro del clúster C2 son las variables con la correlación más alta con respecto a la variable V27.

(i) Se observa que en el clúster C2, considerando el factor de la SBS para la clasificación de los deudores, el 97% de los clientes ha sido clasificado como un deudor normal, el 2% de los clientes ha sido clasificados como un cliente con potencial pérdida, y el resto con igual probabilidad ha sido clasificado como deficiente, dudoso o pérdida.

(ii) Con respecto a la clasificación propia de la entidad financiera sin considerar el factor de la SBS, en este clúster el 97% de los clientes ha sido clasificado como un deudor normal, el 2% ha sido clasificado como un cliente con potencial pérdida, y el resto con igual probabilidad ha sido clasificado como deficiente, dudoso o pérdida.

d) Frecuencia de las variables Monto del crédito otorgado (V2) y Saldo capital de la deuda (V3) en el clúster C2

Tabla 69. Frecuencia de las variables V2 y V3

Frecuencia de V2			Frecuencia de V3		
[8500, 40000]	FREC	PROBAB	[2928.69, 29821.89]	FREC	PROBAB
8500.00	22.00	0.01	2928.69	2.00	0.00
11125.00	514.00	0.34	5169.79	8.00	0.01
13750.00	358.00	0.24	7410.89	124.00	0.08
16375.00	52.00	0.03	9651.99	409.00	0.27
19000.00	323.00	0.21	11893.09	192.00	0.13
21625.00	38.00	0.03	14134.19	260.00	0.17
24250.00	107.00	0.07	16375.29	87.00	0.06
26875.00	20.00	0.01	18616.39	153.00	0.10
29500.00	78.00	0.05	20857.49	142.00	0.09
32125.00	2.00	0.00	23098.59	48.00	0.03
34750.00	10.00	0.01	25339.69	64.00	0.04
37375.00	0	0	27580.79	38.00	0.02
40000.00	2.00	0.00	29821.89	1.00	0.00

Fuente: Elaboración propia

e) Frecuencia de las variables Saldo capital vigente de la operación (V11) y Rendimientos devengados de la operación (V14) en el clúster C2

Tabla 70. Frecuencia de las variables V11 y V14

Frecuencia de V11			Frecuencia de V14		
[0, 28500]	FREC	PROBAB	[0, 6871.68]	FREC	PROBAB
0	8.00	0.01	0	883.00	0.58
2375.00	1.00	0.00	572.64	523.00	0.34
4750.00	8.00	0.01	1145.28	75.00	0.05
7125.00	78.00	0.05	1717.92	24.00	0.02
9500.00	453.00	0.30	2290.56	12.00	0.01
11875.00	196.00	0.13	2863.20	5.00	0.00
14250.00	268.00	0.18	3435.84	2.00	0.00
16625.00	87.00	0.06	4008.48	1.00	0.00
19000.00	229.00	0.15	4581.12	0	0
21375.00	59.00	0.04	5153.76	1.00	0.00

23750.00	59.00	0.04	5726.40	0	0
26125.00	63.00	0.04	6299.04	1.00	0.00
28500.00	19.00	0.01	6871.68	1.00	0.00

Fuente: Elaboración propia

f) Rentabilidad en el clúster C2

- Total aceptados: 1,513 créditos
- Monto de crédito promedio: S/ 16,279.50
- Total monto de créditos otorgados: S/ 24’630,884
- Tasa de interés: 30% anual
- Plazo de pago promedio: 28
- Intereses ganados: S/ 9’912,998
- Rentabilidad en el período de pago: 40.25 %
- Rentabilidad anual: 17.25 %

g) Riesgo en el clúster C2

- Máximo días de atraso de la última cuota pagada: 77 días
- Promedio días de atraso de la última cuota pagada: 1.76 días
- Riesgo: 2.28 %

Estudio del clúster C3

a) Correlaciones en el clúster C3

La Tabla 71 muestra las medidas estadísticas de las variables y su correlación con la variable V27 dentro del clúster C3.

Tabla 71. Estadísticas del clúster C3 (CC con V27)

Variable	Min	Max	STD	Media	Moda	CC
V1	1	2	0.12	1.02	1	0.03
V2	30000	297770	17772.04	43658.99	40000	0.03
V3	22860.75	297770	17366.52	41709.84	41625.02	0.03
V4	8	13	0.93	9.48	9	0.1
V5	0	3	0.33	0.08	0	-0.42
V6	0	1	0.25	0.07	0	-0.54
V7	-419	30	40.83	-28.6	-15	-0.21
V8	**0**	**60**	**9.22**	**3.8**	**0**	**-0.79**
V9	-18	19	1.18	-0.06	0	-0.01
V10	342.91	48000	2500.76	911.37	600	-0.07
V11	22860.75	297770	17366.52	41709.84	41625.02	0.03
V12	0	0	0	0	0	NaN
V13	0	0	0	0	0	NaN
V14	0	19652.81	1236.86	1689.61	1204.62	-0.17
V15	0	2791	144.54	9.06	0	0.01
V16	40547	40847	87.33	40689.56	40751	0.16
V17	1	3	0.4	2.91	3	-0.04
V18	0	120	56.57	67.27	120	-0.09
V19	40869	48084	660.11	42223.57	42700	-0.03
V20	40817	41209	40.22	40875.61	40862	0.22
V21	30	365	46.33	39.65	30	0.04
V22	1	240	22.23	47.82	60	-0.05
V23	0	9	2.13	2.58	0	-0.05
V24	1	1	0	1	1	NaN
V25	1	43	10.14	18.89	15	0.02
V26	12.55	49.36	6.34	22.58	18.02	0.01
V27	-1	1	0.38	0.92	1	1
Clúster C3	3	3	0	3	3	NaN

ZFuente: Elaboración propia

Observaciones sobre el clúster C3:

(i) El clúster C3 ha seleccionado clientes cuyo monto del crédito otorgado (V2) es mayor que en el clúster C2 y mucho mayor que en el clúster C1. Estos créditos se concentran alrededor de la media de 43,659 soles. El

monto del crédito está en el intervalo [30000, 297770] y es reducido el número de clientes que tienen un monto del crédito igual o cercano al mínimo o máximo, esto se explica por el valor de la media. Esta variable ha tenido una participación importante en la formación de los clústeres, esto se explica por el valor de su correlación alta (0.88), en cambio, su correlación con la variable V27 es muy baja, lo que indica que esta variable no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(ii) La variable saldo capital de la deuda (V3), indica que, en el período de repago del préstamo, el saldo del capital varía en el rango de 22,861 hasta 297,770 soles, siendo el promedio de 41,594 soles. Hay un reducido número de clientes que tienen un saldo capital de la deuda igual o cercano a los valores máximos o mínimo, esto se explica por el valor de la media. La mayoría de los clientes tienen un saldo de capital cuyo monto se encuentra alrededor de la media, en el intervalo entre 24,343 y 59,076 soles. El alto grado de correlación de esta variable (0.88) explica porque esta ha sido tomada en cuenta para la formación de los clústeres. En cambio, su correlación con la variable V27 es muy baja, lo que indica que esta variable no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iii) La variable saldo de capital vigente de la operación (V11) varía en el rango de 22,861 a 297,770 soles, su media es 41,710 soles. Es reducido el número de clientes con un saldo de capital vigente de la operación con valores iguales o cercanos al máximo o al mínimo, es decir, la mayoría de los clientes tienen un saldo de capital vigente de la operación alrededor de la media. Esta variable tiene un coeficiente de correlación alto con valor 0.88 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. También es necesario precisar que la correlación de esta

variable con la variable V27 es muy baja, lo que indica que no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(iv) La variable rendimientos devengados de la operación (V14) tiene valores que fluctúan desde un mínimo de cero hasta 19,653 soles. Teniendo en cuenta que la media es de 1,690 soles, se puede afirmar que hay un reducido número de clientes con valores iguales o cercanos al máximo, por lo que, para la mayoría de los clientes, el valor de esta variable se concentra alrededor de la media. La variable V14 tiene un coeficiente de correlación significativamente alto con valor 0.70 lo que explica porque se ha tomado en cuenta en la formación de los clústeres. Es necesario precisar que la correlación con la variable V27 es muy baja, lo que indica que no ha sido tomada en cuenta para la aceptación o rechazo de la solicitud de crédito.

(v) La variable días de atraso de la última cuota pagada (V8) en el clúster C3 varía en el rango de 0 a 60 días. Su media de 3.77 días es ligeramente mayor a la media de esta variable en el clúster C1. Esto indica que la mayoría de los clientes ha tenido un retraso de 4 días para pagar su cuota, solo algunos clientes se han pasado de los 30 días de retraso. Esta variable tiene una correlación baja en la formación de los clústeres, sin embargo, es la variable que tiene la mayor correlación con la variable de aceptación o rechazo de la solicitud de crédito.

b) Frecuencia de las variables Aceptación o rechazo de la solicitud de crédito (V27) y Días de atraso de la última cuota pagada (V8) en el clúster C3

Tabla 72. Frecuencia de las variables V27 y V8

Frecuencia de V27			Frecuencia de V8		
-1 v 1	FREC	PROBABILIDAD	[0, 6871.68]	FREC	PROBAB
-1.00	31.00	0.04	0	636.00	0.77
1.00	792.00	0.96	5.00	74.00	0.09
			10.00	28.00	0.03
			15.00	31.00	0.04
			20.00	15.00	0.02
			25.00	9.00	0.01
			30.00	11.00	0.01
			35.00	10.00	0.01
			40.00	4.00	0.00
			45.00	6.00	0.01
			50.00	1.00	0.00
			55.00	4.00	0.00
			60.00	1.00	0.00

Fuente: Elaboración propia

c) Frecuencia de las variables Clasificación del deudor (V5) y Clasificación del deudor sin considerar alineamiento del sistema (V6) en el clúster C3

Tabla 73. Frecuencia de las variables V5 y V6

Frecuencia de V5			Frecuencia de V6		
[0, 3]	FREC	PROBAB	[0, 1]	FREC	PROBAB
0	766.00	0.93	0	769.00	0.93
1.00	51.00	0.06	1.00	54.00	0.07
2.00	2.00	0.00			
3.00	4.00	0.00			

Fuente: Elaboración propia

Comentarios sobre las frecuencias de las variables V5 y V6 en el clúster C3:

Las variables V5 y V6, después de la variable V8, dentro del clúster C2 son las variables con la correlación más alta con respecto a la variable V27.

(i) Se observa que en el clúster C3, considerando el factor de la SBS para la clasificación de los deudores, el 93% de los clientes ha sido clasificado como un deudor normal, el 3% de los clientes ha sido clasificados como un cliente con potencial pérdida, y el resto ha sido clasificado como deficiente o dudoso. Nadie ha sido clasificado como pérdida.

(ii) Con respecto a la clasificación propia de la entidad financiera sin considerar el factor de la SBS, en este clúster los clientes sólo han sido clasificados en dos categorías: el 93% de los clientes ha sido clasificado como un deudor normal y el 7% ha sido clasificado como un cliente con potencial pérdida. Nadie ha sido clasificado como deficiente, dudoso o pérdida.

d) Frecuencia de las variables Monto del crédito otorgado (V2) y Saldo capital de la deuda (V3) en el clúster C3

Tabla 74. Frecuencia de las variables V2 y V3

Frecuencia de V2			Frecuencia de V3		
[30000, 297770]	FREC	PROBAB	[22860.75, 297770]	FREC	PROBAB
30000.00	317.00	0.39	22860.75	167.00	0.20
52314.17	465.00	0.57	45769.85	611.00	0.74
74628.33	19.00	0.02	68678.96	23.00	0.03
96942.50	11.00	0.01	91588.06	12.00	0.01
119256.67	7.00	0.01	114497.17	6.00	0.01
141570.83	1.00	0.00	137406.27	1.00	0.00
163885.00	1.00	0.00	160315.38	1.00	0.00
186199.17	0	0	183224.48	0	0

208513.33	0	0	206133.58	0	0
230827.50	0	0	229042.69	0	0
253141.67	0	0	251951.79	0	0
275455.83	1.00	0.00	274860.90	1.00	0.00
297770.00	1.00	0.00	297770.00	1.00	0.00

Fuente: Elaboración propia

e) Frecuencia de las variables Saldo capital vigente de la operación (V11) y Rendimientos devengados de la operación (V14) en el clúster C3

Tabla 75. Frecuencia de las variables V11 y V14

Frecuencia de V11			Frecuencia de V14		
[22860.75, 297770]	FREC	PROBAB	[0, 19652.81]	FREC	PROBAB
22860.75	167.00	0.21	0	249.00	0.31
45769.85	611.00	0.74	1637.73	354.00	0.43
68678.96	23.00	0.03	3275.47	211.00	0.25
91588.06	12.00	0.01	4913.20	7.00	0.01
114497.17	6.00	0.01	6550.94	1.00	0.00
137406.27	1.00	0.00	8188.67	0	0
160315.38	1.00	0.00	9826.41	1.00	0.00
183224.48	0	0	11464.14	0	0
206133.58	0	0	13101.87	0	0
229042.69	0	0	14739.61	0	0
251951.79	0	0	16377.34	0	0
274860.90	1.00	0.00	18015.08	0	0
297770.00	1.00	0.00	19652.81	1.00	0.00

Fuente: Elaboración propia

f) Rentabilidad en el clúster C3

- Total aceptados: 792 créditos
- Monto de crédito promedio: S/ 43,658.99

- Total monto de créditos otorgados: S/ 34’577,920
- Tasa de interés: 30% anual
- Plazo de pago promedio: 48 meses
- Intereses ganados: S/ 25’182,674
- Rentabilidad en el período de pago:
- Rentabilidad anual: 18.21%

g) Riesgo en el clúster C3

- Máximo días de atraso de la última cuota pagada: 60 días
- Promedio días de atraso de la última cuota pagada: 3.8 días
- Riesgo: 6.33 %

Conclusiones de la prueba

Debido a que las variables V2, V3, V11 y V14 tienen la correlación más alta en la formación de los clústeres, luego de su estudio, se tienen las siguientes conclusiones:

(i) El cambio a la topología Hextop y a la métrica Linkdist no ha afectado significativamente los resultados obtenidos con la topología Gridtop y la métrica Dist. Esta afirmación se sustenta en que se ha obtenido las mismas medidas estadísticas y los coeficientes de correlación en la formación de los 3 clústeres, y en cada clúster las medidas estadísticas y coeficientes de correlación tienen resultados iguales o muy similares.

(ii) Los clientes agrupados en el clúster C1, presentan un monto del crédito (V2) que está en el intervalo [300, 20000] con una media de 2,325 soles, que se considera como montos de crédito bajo. En el clúster C2 están los

clientes con monto del crédito moderadamente alto, pues estos montos se encuentran en el intervalo [8500, 40000] y se dispersan alrededor de la media de 16,280 soles. Es necesario indicar que, de la intersección de los dos intervalos, se obtiene un conjunto que contiene a clientes que pueden estar en el clúster C1 o en el clúster C2. En el clúster C3 están los clientes con monto del crédito que están en el intervalo [30000, 297770] con una media de 43,659 soles, que se consideran montos de crédito alto. También se debe precisar que, al intersectar los intervalos de los clústeres C2 y C3, se obtiene un conjunto que contiene a clientes que pueden pertenecer al clúster C2 o al clúster C3.

(iii) En el clúster C1 se agrupan los clientes cuyo saldo capital de la deuda (V3) está en el intervalo [16, 8581] y que tienen una media de 1723 soles, considerados como que tienen saldo capital bajo. En el clúster C2 están los clientes con saldo capital de la deuda moderadamente alto, porque se encuentran en el intervalo [2929, 29822] y se distribuyen alrededor de la media de 14,373 soles. Hasta aquí se debe indicar que, de la intersección de los dos intervalos, se obtiene un conjunto que contiene a clientes que pueden estar en el clúster C1 o en el clúster C2. En el clúster C3 están los clientes con saldo capital de la deuda alto, pues se agrupan en el intervalo [22861, 297770] y se dispersan alrededor de la media de 41,710 soles. Es necesario precisar que, al intersectar los intervalos de los clústeres C2 y C3, se obtiene un conjunto que contiene a clientes que pueden pertenecer al clúster C2 o al clúster C3.

(iv) Con respecto a la variable saldo de capital vigente de la operación (V11), en el clúster C1 están agrupados los clientes con valores de esta variable que se encuentran en el intervalo [0, 8294] y que se dispersan alrededor de la media de 1,685 soles, considerados como saldos de capital bajo. En el clúster C2 se agrupan los clientes con valores de esta variable que están en el intervalo [0, 28500] con una media de 14,259 soles,

considerado como saldos de capital moderadamente alto. En el clúster C3 se agrupan los clientes con saldos de capital que están en el intervalo [22861, 297770] con una media de 41,710 soles, y se considera como saldos de capital vigente altos. También se indica que, al intersectar los intervalos de los clústeres C2 y C3, se obtiene un conjunto que contiene a clientes que pueden pertenecer al clúster C2 o al clúster C3.

(v) Con referencia a la variable rendimientos devengados de la operación (V14), se aprecia que en el clúster C1 está agrupados los clientes cuyos rendimientos devengados se encuentran en el intervalo [0, 1782] con una media de 46 soles, considerado bajo. En el clúster C2 están agrupados los clientes cuyo valor de esta variable se encuentra en el intervalo [0, 6872] y están dispersos alrededor de la media de 369 soles, valor que es mayor que en el clúster C1. En el clúster C3 están agrupados los clientes cuyo valor de esta variable se encuentra en el intervalo [0, 19653] y se dispersan alrededor de la media de 1,690 soles, valor que es mayor que en el clúster C2.

(vi) En el clúster C1, la variable días de atraso de la última cuota pagada (V8) está en el intervalo [0, 162] y tiene una media de 2.96 días. Esto indica que la mayoría de los clientes ha pagado con un retraso de alrededor de 3 días, y muy pocos clientes han pagado con un retraso igual o cercano a 162 días. En el clúster C2, esta variable fluctúa en el rango de 0 a 77 días, su media de 1.76 días es menor a la media de esta variable en el clúster C1. Esto indica que la mayoría de los clientes ha tenido un retraso de 2 días para pagar su cuota, solo algunos clientes se han pasado de los 30 días. En el clúster C3 varía en el rango de 0 a 60 días. Su media de 3.77 días es ligeramente mayor a la media de esta variable en el clúster C1. Esto indica que la mayoría de los clientes ha tenido un retraso de 4 días para pagar su cuota, solo algunos clientes han tenido un retraso mayor a 30 días. Esta variable tiene una correlación baja en la formación

de los clústeres, pero es la que tiene la más alta correlación con la variable de aceptación o rechazo de la solicitud de crédito.

(vii) La rentabilidad anual en el clúster C1 es 16.98%, en el clúster C2 es 17.25%, y en el clúster C3 cuyo valor es 18.21%. El riesgo en el clúster C1 es 1.83%, en el clúster C2 es 2.28% y en el clúster C3 es 6.33%

(viii) Se concluye que, el grupo más numeroso es el clúster C1 y está conformado por clientes cuyos valores de estas cuatro variables son bajos; en este clúster la rentabilidad y el riesgo son bajos. En el clúster C2 están agrupados los clientes con valores de estas variables moderadamente altos; la rentabilidad y el riesgo son ligeramente mayor que en el clúster C1. En el clúster C3 están agrupados los clientes con valores altos en estas cuatro variables; la rentabilidad es ligeramente mayor que en el clúster C2, en cambio el riesgo es significativamente mayor que en el clúster C2.

(ix) No hay conjuntos disjuntos. Al intersectar dos clústeres contiguos, se obtiene un conjunto que contiene clientes que están en uno de los dos clústeres intersectados. La Tabla 76 resume las características de los registros en los clústeres.

Tabla 76. Red SOM de tres neuronas con topología Hextop y métrica Linkdist

Clúster	Tamaño	Aceptados	Rechazados	Valores de variables	Rentabilidad	Riesgo
C1	13,218	12,951	267	Bajo	16.98%	1.83%
C2	1,528	1,513	15	Medio	17.25%	2.28%
C3	823	792	31	Alto	18.21%	6.33%

Fuente: Elaboración propia

4.2.3. Máquinas con Soporte Vectorial (MSV)

Se analiza el resultado de la MSV con núcleo lineal para separar los prestatarios en dos grupos y determinar los vectores soporte.

(I) Análisis de los resultados con la BD completa

Se analizan los resultados de los vectores soporte obtenidos por la MSV.

Comparación de F(-) y F(+)

La Tabla 77 contiene las medidas estadísticas y el coeficiente de correlación con la categoría de los registros que están en las fronteras.

Tabla 77. Medidas estadísticas de los registros en las fronteras F(-) y F(+)

Variable	Min	Max	Media	Moda	STD	CC
V1	1	2	1.02	1	0.16	0.21
V2	400	297770	19031.97	500	47592.97	0.19
V3	16.24	297770	17898.5	2000	47650.23	0.20
V4	9	12	10.51	12	1.27	0.13
V5	0	4	1.41	1	1.36	-0.05
V6	0	4	1.34	1	1.35	-0.12
V7	-360	170	23.83	14	81.25	-0.18
V8	0	61	25.78	0	19.62	-0.93
V9	0	31	3.02	0	6.57	-0.2
V10	2.05	48000	3293.82	30	9196.65	0.17
V11	0	297770	15349.9	0	47927.39	0.19
V12	0	25104	1333.44	0	5128.23	0.05
V13	0	29822	1215.17	0	5542.58	0.02
V14	0	6871.68	597.62	0	1508.49	0.14
V15	0	5903	360.12	0	1096.3	0.16

V16	40557	40847	40658.85	40611	88.86	0.38
V17	1	5	2.9	3	0.62	-0.21
V18	0	120	7.41	0	26.37	-0.15
V19	40714	45818	41409.85	40714	895.06	0.05
V20	40677	41207	40827.41	40831	84.23	0.24
V21	1	360	47.59	30	73.15	0.32
V22	1	168	24.41	6	29.38	0.02
V23	0	31	4.12	0	4.98	-0.04
V24	1	1	1	1	0	NaN
V25	1	43	20.07	35	13.12	-0.08
V26	14.03	293.79	48.92	58.27	44.01	0.06
Cat	-1	1	-0.27	-1	0.98	1

Fuente: Elaboración propia

Observaciones:

(a) La variable Días de atraso de la última cuota pagada (V8) es la que tiene el coeficiente de correlación significativamente alto con valor -0.93

(b) Las variables que tienen un coeficiente de correlación bajo pero significativo son Monto del crédito otorgado (V2), Saldo capital de la deuda (V3), Saldo capital vigente de la operación (V11), Periodicidad de las cuotas (V21), Días de atraso al cierre del mes (V7) y Clasificación del deudor sin considerar alineamiento con el sistema (V6). Estas variables se tendrán en cuenta para hacer un análisis estadístico de los vectores soporte. Es necesario indicar que las variables Fecha de desembolso (V16) y Fecha de vencimiento puntual de la operación (V20) también muestran un coeficiente de correlación bajo, pero no se analizarán por ser variables tipo fecha.

Estadísticas de los vectores soporte F(+) (Aceptados)

La Tabla 78 muestra las medidas estadísticas de los vectores soporte de los créditos Aceptados.

Tabla 78. Estadísticas de los vectores soporte F(+) Aceptados

Variable	Min	Max	Media	Moda	STD
V1	1	2	1.07	1	0.26
V2	400	297770	30617.33	500	76871.14
V3	16.24	297770	30147.2	2000	76944.08
V4	9	12	10.73	12	1.28
V5	0	4	1.33	0	1.76
V6	0	4	1.13	0	1.73
V7	-360	170	4.6	-10	125.66
V8	0	26	2.13	0	6.69
V9	0	14	1.33	0	3.62
V10	2.05	48000	5360.56	30	12909.75
V11	0	297770	27115.37	0	77671.18
V12	0	25104	1698.53	0	6475.31
V13	0	20000	1333.33	0	5163.98
V14	0	6871.68	869.6	0	2287.18
V15	0	5903	594	0	1559.49
V16	40557	40847	40703.2	40847	120.05
V17	1	5	2.73	3	1.03
V18	0	30	2.13	0	7.73
V19	40714	45818	41471.87	40714	1309.74
V20	40677	41207	40854	40853	128.96
V21	1	360	78.07	30	117.13
V22	1	168	25.27	1	43.33
V23	0	31	3.87	0	7.85
V24	1	1	1	1	0
V25	1	37	18.73	1	14.27
V26	14.03	293.79	52.25	18.16	69.75
Cat	1	1	1	1	0

Fuente: Elaboración propia

La Tabla 79 muestra las medidas estadísticas de los vectores soporte de los créditos Rechazados.

Tabla 79. Estadísticas de los vectores soporte F(-) Rechazados

Variable	Min	Max	Media	Moda	STD
V1	1	1	1	1	0
V2	500	45418.18	12348.11	500	13711.86
V3	93.11	44387.21	10831.94	93.11	13247.66
V4	9	12	10.38	10	1.27
V5	0	4	1.46	1	1.1
V6	0	4	1.46	1	1.1
V7	0	134	34.92	14	37.01
V8	31	61	39.42	32	7.88
V9	0	31	4	0	7.69
V10	13.24	29821.89	2101.47	13.24	6160.63
V11	0	44387.21	8562.13	0	12880.52
V12	0	21091	1122.81	0	4297.82
V13	0	29822	1147	0	5848.58
V14	0	3138.56	440.7	0	800.88
V15	0	3255	225.19	0	713.67
V16	40560	40777	40633.27	40617	52
V17	3	3	3	3	0
V18	0	120	10.46	0	32.45
V19	40773	42582	41374.08	40773	563.5
V20	40713	40847	40812.08	40831	37.01
V21	30	30	30	30	0
V22	6	60	23.92	6	18.13
V23	0	8	4.27	6	2.25
V24	1	1	1	1	0
V25	1	43	20.85	13	12.64
V26	18.02	79.59	47	58.27	19.08
Cat	-1	-1	-1	-1	0

Fuente: Elaboración propia

La Tabla 80 es una tabla comparativa de las medidas estadísticas de ambos vectores Aceptados y Rechazados.

Tabla 80. Medidas estadísticas de ambos vectores soporte

Aceptados (F+)						Rechazados (F-)				
Var	**Min**	**Max**	**Media**	**Moda**	**STD**	**Min**	**Max**	**Media**	**Moda**	**STD**
V1	1	2	1.07	1	0.26	1	1	1	1	0
V2	400	297770	30617.3	500	76871.1	500	45418	12348	500	13711.9
V3	16.24	297770	30147.2	2000	76944.1	93.11	44387	10832	93.11	13247.7
V4	9	12	10.73	12	1.28	9	12	10.38	10	1.27
V5	0	4	1.33	0	1.76	0	4	1.46	1	1.1
V6	0	4	1.13	0	1.73	0	4	1.46	1	1.1
V7	-360	170	4.6	-10	125.66	0	134	34.92	14	37.01
V8	0	26	2.13	0	6.69	31	61	39.42	32	7.88
V9	0	14	1.33	0	3.62	0	31	4	0	7.69
V10	2.05	48000	5360.56	30	12909.8	13.24	29822	2101.5	13.24	6160.63
V11	0	297770	27115.4	0	77671.2	0	44387	8562.1	0	12880.5
V12	0	25104	1698.53	0	6475.31	0	21091	1122.8	0	4297.82
V13	0	20000	1333.33	0	5163.98	0	29822	1147	0	5848.58
V14	0	6871.7	869.6	0	2287.18	0	3139	440.7	0	800.88
V15	0	5903	594	0	1559.49	0	3255	225.19	0	713.67
V16	40557	40847	40703.2	40847	120.05	40560	40777	40633	40617	52
V17	1	5	2.73	3	1.03	3	3	3	3	0
V18	0	30	2.13	0	7.73	0	120	10.46	0	32.45
V19	40714	45818	41471.9	40714	1309.74	40773	42582	41374	40773	563.5
V20	40677	41207	40854	40853	128.96	40713	40847	40812	40831	37.01
V21	1	360	78.07	30	117.13	30	30	30	30	0
V22	1	168	25.27	1	43.33	6	60	23.92	6	18.13
V23	0	31	3.87	0	7.85	0	8	4.27	6	2.25
V24	1	1	1	1	0	1	1	1	1	0
V25	1	37	18.73	1	14.27	1	43	20.85	13	12.64
V26	14.03	293.79	52.25	18.16	69.75	18.02	79.59	47	58.27	19.08
Cat	1	1	1	1	0	-1	-1	-1	-1	0

Fuente: Elaboración propia

Observaciones:

(a) Con respecto a la variable Días de atraso de la última cuota pagada (V8), los clientes que están en la frontera de los Aceptados han tenido un retraso que fluctúa en el intervalo [0, 26] días, en promedio el retraso en

el pago ha sido de 2 días. En la frontera de los clientes Rechazados, el retraso en el pago fluctúa en el intervalo [31, 61] días, y en promedio el retraso en el pago es de 39 días. Claramente se aprecia que, para esta variable, los conjuntos son disjuntos.

(b) La variable Monto del crédito otorgado (V2) en los vectores frontera de los clientes Aceptados, tiene valores que fluctúan en el intervalo [400, 297770] y su promedio es de 30,617 soles. En la frontera de los clientes Rechazados, el monto del crédito fluctúa en el intervalo [500, 45418] y su media es de 12,348 soles. Se observa que el intervalo de variación del crédito en los Aceptados es de mayor extensión, de la misma manera, el promedio de crédito de los Aceptados es mayor que los Rechazados.

(c) Con respecto a la variable Saldo capital de la deuda (V3), en los vectores frontera de los créditos Aceptados, fluctúa en el intervalo [16, 297770] y su media es de 30,147 soles. En los vectores frontera de los Rechazados, el saldo capital varía en el intervalo [93, 44387] y su media es de 10,832 soles. Con esta variable también se aprecia que tanto el intervalo como el promedio son mayores en los vectores frontera de los créditos Aceptados.

(d) La variable Clasificación del deudor sin considerar alineamiento con el sistema (V6) tiene el mismo mínimo y máximo en ambos vectores frontera, sin embargo, en los créditos Aceptados el 67% ha sido clasificado como cliente normal y nadie ha sido clasificado como cliente con potencial pérdida (CPP), en cambio, en los créditos Rechazados, sólo el 12% ha sido clasificado como cliente normal y el 58% ha sido clasificado como CPP. En la Tabla 81 se puede apreciar la Clasificación del deudor en ambos vectores soporte.

Tabla 81. Frecuencia de la variable V6 en Aceptados y Rechazados

Frecuencia de V6 Aceptados			Frecuencia de V6 Rechazados		
[0, 4]	FREC	PROBAB	[0, 4]	FREC	PROBAB
0	10.00	0.67	0	3.00	0.12
1.00	0	0	1.00	15.00	0.58
2.00	1.00	0.07	2.00	3.00	0.12
3.00	1.00	0.07	3.00	3.00	0.12
4.00	3.00	0.20	4.00	2.00	0.08

Fuente: Elaboración propia

(e) Con respecto a la variable Días de atraso al cierre del mes (V7), en la frontera de los créditos aceptados, el promedio es de 5 días, y en los Rechazados el promedio es de 35 días.

(f) La variable Saldo capital vigente de la operación (V11), en los vectores soporte de los créditos Aceptados, tiene valores que fluctúa en el intervalo [0, 297770] con un promedio de 27,115 soles. En los vectores soporte de los créditos Rechazados, esta variable fluctúa en el intervalo [0, 44387] y su promedio es de 8,562 soles.

(g) La variable Periodicidad de cuotas (V21) en los vectores frontera de los créditos Aceptados la periodicidad promedio de las cuotas es de 78 días, y para los créditos Rechazados todos tienen una periodicidad de 30 días.

Conclusiones de la prueba MSV con la base de datos completa:

(i) En cuanto a las variables Días de atraso de la última cuota pagada (V8) y Días de atraso al cierre del mes (V7), en los vectores frontera de los créditos Aceptados el promedio es menor que en los vectores frontera de los créditos Rechazados.

(ii) Con respecto a las variables Monto del crédito otorgado (V2), Saldo capital de la deuda (V3) y Saldo capital vigente de la operación (V11), en

los vectores frontera de los créditos Aceptados el valor promedio de esta variable es mayor que en los créditos Rechazados.

(iii) Los valores de la variable Clasificación del deudor sin considerar alineamiento con el sistema (V6), muestran que en los vectores frontera de los créditos Aceptados un gran porcentaje de los clientes ha sido clasificado como normal, en cambio en los créditos Rechazados, el mayor porcentaje ha sido clasificado como cliente con potencial pérdida.

(iv) La variable Periodicidad de cuota (V21), en los vectores frontera de los créditos Aceptados tiene un valor promedio de 78 días, en comparación con los créditos Rechazados en que todos tienen una periodicidad de 30 días.

(II) Análisis de los resultados con la BD reducida

Se analizan los resultados obtenidos con la MSV excluyendo la variable V8 de la base de datos.

Comparación de F(-) y F(+)

La Tabla 82 muestra las medidas estadísticas y el coeficiente de correlación con la categoría de los registros que están en las fronteras.

Tabla 82. Estadísticas de los vectores frontera F(+) y F(-)

Variable	Min	Max	Media	Moda	STD	CC
V1	1	1	1	1	0	NaN
V2	400	99900	21583.67	1000	23944.84	0.20
V3	16.24	99900	18297.42	16.24	23676.83	0.13
V4	9	13	10.54	9	1.43	0.13

V5	0	4	1.14	0	1.3	-0.22
V6	0	4	1.04	0	1.26	-0.32
V7	-167	151	15.18	6	65.44	-0.37
V9	0	33	4	0	9.27	-0.31
V10	1.55	48000	2944.47	1.55	9118.53	0.13
V11	0	99900	16063.64	0	24214.07	0.14
V12	0	26527	2233.79	0	6858.57	-0.06
V13	0	0	0	0	0	NaN
V14	0	19652.81	1285.04	0	3858.87	0.19
V15	0	1799	280.29	0	606.29	-0.11
V16	40547	40847	40661.89	40611	75.82	0
V17	1	3	2.93	3	0.38	-0.19
V18	0	120	9.71	0	27.99	0.01
V19	40714	47893	41920.93	40714	1935.25	0.26
V20	40696	41014	40838.04	40831	69.92	0.43
V21	30	360	53.57	30	86.55	0.28
V22	1	240	40.5	12	64.62	0.25
V23	0	11	3.25	0	2.74	-0.07
V24	1	1	1	1	0	NaN
V25	1	42	21.25	13	12.47	0.04
V26	12.01	79.59	39.13	34.49	19.07	-0.23
Categoría	-1	1	0	-1	1.02	1

Fuente: Elaboración propia

Observaciones:

(a) La variable que tiene la mayor correlación es Fecha de vencimiento puntual de la operación (V20). Se trata de una variable fecha, que no merece mayor comentario.

(b) Las variables que tienen un coeficiente de correlación bajo pero significativo son Días de atraso al cierre del mes (V7), Clasificación de deudor sin considerar alineamiento con el sistema (V6) y Promedio de días de atraso en el pago de cuotas en los últimos 6 meses (V9). Se hará un análisis estadístico de estas variables en los vectores soporte.

Estadísticas de los vectores soporte F(+) (Aceptados)

La Tabla 83 muestra las medidas estadísticas de los vectores soporte de los créditos Aceptados.

Tabla 83. Estadísticas de los vectores soporte F(+) Aceptados

Variable	Min	Max	Media	Moda	STD
V1	1	1	1	1	0
V2	600	99900	25138.77	20000	25241.09
V3	16.24	99900	22325.47	16.24	25330.63
V4	9	13	10.5	9	1.56
V5	0	4	1.14	0	1.46
V6	0	4	1.14	0	1.46
V7	-167	151	10.43	52	83.29
V9	0	33	5.57	0	11.8
V10	1.55	8581.21	1766.56	1.55	2693.71
V11	0	99900	19811.47	0	26419.61
V12	0	25104	2514	0	6890.85
V13	0	0	0	0	0
V14	0	19652.81	2323.16	0	5332.95
V15	0	1799	341.36	0	681.19
V16	40547	40640	40603.57	40611	26.73
V17	1	3	2.86	3	0.53
V18	0	120	17.36	0	38.42
V19	40714	47893	42296.86	40714	2418.92
V20	40696	41014	40849	40696	89.37
V21	30	360	77.14	30	119.84
V22	1	240	51.64	12	81.67
V23	0	11	4.5	5	2.95
V24	1	1	1	1	0
V25	1	37	17.21	13	11.66
V26	12.01	69.59	33.47	12.01	17
Cat	1	1	1	1	0

Fuente: Elaboración propia

Estadísticas de los vectores soporte F(-) (Rechazados)

La Tabla 84 muestra las medidas estadísticas de los vectores soporte de los créditos Rechazados.

Tabla 84. Estadísticas de los vectores soporte F(-) Rechazados

Variable	Min	Max	Media	Moda	STD
V1	1	1	1	1	0
V2	400	80000	1000	18028.57	22945.17
V3	177.02	80000	177.02	14269.38	22084.37
V4	9	12	12	10.57	1.34
V5	0	3	0	1.14	1.17
V6	0	3	0	0.93	1.07
V7	-32	113	6	19.93	43.68
V9	0	18	0	2.43	5.83
V10	2.66	48000	2.66	4122.38	12745.46
V11	0	80000	0	12315.81	22124.51
V12	0	26527	0	1953.57	7074.28
V13	0	0	0	0	0
V14	0	1242	0	246.91	407.11
V15	0	1787	0	219.21	539.83
V16	40648	40847	40703	40720.21	62.43
V17	3	3	3	3	0
V18	0	17	0	2.07	5.06
V19	40826	45818	40826	41545	1273.94
V20	40734	40879	40841	40827.07	43.68
V21	30	30	30	30	0
V22	4	168	12	29.36	41.65
V23	0	5	0	2	1.88
V24	1	1	1	1	0
V25	8	42	13	25.29	12.33
V26	18.16	79.59	58.27	44.79	19.93
Cat	-1	-1	-1	-1	0

Fuente: Elaboración propia

La Tabla 85 es una tabla comparativa de las medidas estadísticas de ambos vectores Aceptados y Rechazados.

Tabla 85. Medidas estadísticas de ambos vectores soporte

Aceptados F(+)						Rechazados F(-)				
Var	**Min**	**Max**	**Media**	**Moda**	**STD**	**Min**	**Max**	**Media**	**Moda**	**STD**
V1	1	1	1	1	0	1	1	1	1	0
V2	600	99900	25138.8	20000	25241.1	400	80000	1000	18029	22945.2
V3	16.24	99900	22325.5	16.24	25330.6	177	80000	177	14269	22084.4
V4	9	13	10.5	9	1.56	9	12	12	10.57	1.34
V5	0	4	1.14	0	1.46	0	3	0	1.14	1.17
V6	0	4	1.14	0	1.46	0	3	0	0.93	1.07
V7	-167	151	10.43	52	83.29	-32	113	6	19.93	43.68
V9	0	33	5.57	0	11.8	0	18	0	2.43	5.83
V10	1.55	8581.2	1766.56	1.55	2693.71	2.66	48000	2.66	4122.4	12745.5
V11	0	99900	19811.5	0	26419.6	0	80000	0	12316	22124.5
V12	0	25104	2514	0	6890.85	0	26527	0	1953.6	7074.28
V13	0	0	0	0	0	0	0	0	0	0
V14	0	19653	2323.16	0	5332.95	0	1242	0	246.91	407.11
V15	0	1799	341.36	0	681.19	0	1787	0	219.21	539.83
V16	40547	40640	40603.6	40611	26.73	40648	40847	40703	40720	62.43
V17	1	3	2.86	3	0.53	3	3	3	3	0
V18	0	120	17.36	0	38.42	0	17	0	2.07	5.06
V19	40714	47893	42296.9	40714	2418.92	40826	45818	40826	41545	1273.94
V20	40696	41014	40849	40696	89.37	40734	40879	40841	40827	43.68
V21	30	360	77.14	30	119.84	30	30	30	30	0
V22	1	240	51.64	12	81.67	4	168	12	29.36	41.65
V23	0	11	4.5	5	2.95	0	5	0	2	1.88
V24	1	1	1	1	0	1	1	1	1	0
V25	1	37	17.21	13	11.66	8	42	13	25.29	12.33
V26	12.01	69.59	33.47	12.01	17	18.16	79.59	58.27	44.79	19.93
Cat	1	1	1	1	0	-1	-1	-1	-1	0

Fuente: Elaboración propia

Observaciones:

(a) Con respecto a la variable Días de atraso al cierre del mes (V7), los clientes que están en la frontera de los Aceptados han tenido en promedio un atraso de 10 días al cierre del mes, en cambio los clientes que están en la frontera de los Rechazados han tenido en promedio un atraso de 6 días.

(b) La variable Clasificación del deudor sin considerar alineamiento con el sistema(V6), en la Tabla 86 se aprecia que, en los vectores frontera de Aceptados, el 50% de los clientes ha sido clasificado como un cliente normal y un 14% han sido clasificado como cliente con pérdida potencial; en cambio en los vectores frontera de los Rechazados, un porcentaje menor, el 43% de los clientes ha sido clasificado como cliente normal, y el 36% han sido clasificado como cliente con pérdida potencial.

Tabla 86. Frecuencia de la variable V6 en Aceptados y Rechazados

Frecuencia de V6 Aceptados			**Frecuencia de V6 Rechazados**		
[0, 4]	FREC	PROBAB	[0,3]	FREC	PROBAB
0	7.00	0.50	0	6.00	0.43
1.00	2.00	0.14	0.75	5.00	0.36
2.00	3.00	0.21	1.50	0	0
3.00	0	0	2.25	1.00	0.07
4.00	2.00	0.14	3.00	2.00	0.14

Fuente: Elaboración propia

(c) Con respecto a la variable Promedio de días de atraso en el pago de cuotas en los últimos 6 meses (V9), en los vectores frontera de los créditos Aceptados, esta variable tiene un promedio de 6 días, en cambio, en la frontera de los créditos Rechazados el promedio es 0 días.

Conclusiones de la prueba MSV con la base de datos con 26 variables

(i) Al excluir la variable Días de atraso de la última cuota pagada (V8) de la Base de datos, las variables que tienen la mayor correlación son otras. La

única variable que se mantiene con un coeficiente de correlación significativo es la Clasificación del deudor sin considerar alineamiento con el sistema (V6).

(ii) Aun cuando el número de vectores soporte es reducido, la variable que explica mejor cómo has sido separados los Aceptados de los Rechazados, es la variable V6.

(iii) Hay muy pocos clientes que están en la frontera que podrían representar cierta incertidumbre. Con la MSV los clientes están bien separados, los aceptados y los rechazados.

4.3. CONTRASTACIÓN DE LA HIPÓTESIS

Para realizar el contraste de la hipótesis se utilizan los resultados de las pruebas hechas con las técnicas de aprendizaje de máquina.

4.3.1. Hipótesis General

Los resultados obtenidos de las pruebas con Máquinas de Aprendizaje demuestran que se pueden usar estas técnicas para predecir el comportamiento crediticio de los solicitantes de microcrédito, agruparlos convenientemente en clústeres, y clasificar los registros como aceptados o rechazados.

4.3.2. Hipótesis Específicas

Hipótesis Específica 1

Los resultados obtenidos de las pruebas con las Redes Neuronales Artificiales Backpropagation con determinada arquitectura, demuestran que con esta técnica se predice el comportamiento crediticio de los solicitantes de microcrédito, con alta precisión.

Hipótesis Específica 2

Los resultados obtenidos de las pruebas con las Redes Self Organizing Maps con determinada topología y métrica, demuestran que con esta técnica se agrupa convenientemente a los solicitantes en clústeres.

Hipótesis Específica 3

Los resultados obtenidos de las pruebas con las Máquinas con Soporte Vectorial demuestran que con esta técnica se separan a los solicitantes de microcrédito en aceptados o rechazados.

CONCLUSIONES

(1) Se han hecho diversas pruebas con las Redes Neuronales Artificiales Backpropagation, con diferentes arquitecturas, teniendo como entrada la base de datos completa, y el mejor resultado obtenido es una precisión de 0.97682. Dado que se ha encontrado que la variable Aceptación o rechazo del crédito (V27) depende directamente de la variable Días de atraso de la última cuota pagada (V8), al reemplazar la variable V27 por la variable V8, la precisión de la RNA baja a 0.77389 que sigue siendo una buena precisión. Finalmente, si se elimina de la base de datos la variable V8, la precisión de la RNA baja a 0.66897, con lo que se corrobora que la variable V8 es la más realista, y que la variable V27 es artificial.

(2) La red SOM de dos neuronas con topología Gridtop y métrica Dist ha agrupado en el clúster C1 a los clientes con montos de crédito altos cuya rentabilidad y riesgo es mayor, y en el clúster C2 ha agrupado a los clientes con montos bajos, que a su vez tienen una rentabilidad y riesgo menor.

(3) La red SOM de dos neuronas con topología Hextop y métrica Linkdist ha agrupado en el clúster C1 a los clientes con montos de crédito altos cuya rentabilidad y riesgo es mayor, y en el clúster C2 ha agrupado a los clientes con montos de crédito bajos, que a su vez a su vez tienen una rentabilidad y riesgo menor.

(4) En las redes SOM de dos neuronas, el cambio a la topología Hextop y métrica Linkdist no ha afectado significativamente los resultados obtenidos con la topología Gridtop y métrica Dist, porque no ha habido cambios sustanciales. Esta afirmación se sustenta en que se ha obtenido

las mismas medidas estadísticas y los coeficientes de correlación en la formación de los 2 clústeres, y en cada clúster las medidas estadísticas y coeficientes de correlación tienen resultados iguales o similares. Las diferencias encontradas se deben a que ha habido un desplazamiento de clientes, es decir, algunos clientes se han incorporado a un clúster y otros han salido para incorporarse a otro clúster.

(5) La red SOM de tres neuronas con topología Gridtop y métrica Dist ha agrupado en el clúster C1 al conjunto más numeroso de clientes, con montos de crédito bajos, y cuya rentabilidad y riesgo también son bajos. En el clúster C2 están agrupados los clientes con montos de crédito moderadamente altos, y cuya rentabilidad y riesgo son mayores que los del clúster C1. En el clúster C3 están agrupados los clientes con montos de crédito altos, la rentabilidad es ligeramente mayor que en el clúster C2, pero el riesgo tiene un valor significativamente más alto.

(6) La red SOM de tres neuronas con topología Hextop y métrica Linkdist ha agrupado en el clúster C1 al conjunto más numeroso de clientes, con montos de crédito bajos; en este clúster la rentabilidad y el riesgo son bajos. En el clúster C2 están agrupados los clientes con montos de crédito moderadamente altos; la rentabilidad y el riesgo son ligeramente mayor que en el clúster C1. En el clúster C3 están agrupados los clientes con montos de crédito altos; la rentabilidad es ligeramente mayor que en el clúster C2, en cambio el riesgo es significativamente mayor que en el clúster C2.

(7) El cambio a la topología Hextop y a la métrica Linkdist no ha afectado significativamente los resultados obtenidos con la topología Gridtop y la métrica Dist, porque no ha habido cambios sustanciales. Esta afirmación se sustenta en que se ha obtenido las mismas medidas estadísticas y los coeficientes de correlación en la formación de los 3 clústeres, y en cada clúster las medidas estadísticas y coeficientes de correlación tienen

resultados iguales o similares. Las diferencias encontradas se deben a que ha habido un desplazamiento de clientes, es decir, algunos clientes se han incorporado a un clúster y otros han salido para incorporarse a otro clúster contiguo.

(8) La robustez de esta herramienta ha quedado demostrada porque los cambios en la topología y la métrica no han afectado significativamente los resultados del agrupamiento.

(9) La Máquina con Soporte Vectorial (MSV) ha separado con éxito a los clientes en dos grupos: aceptados y rechazados. Con esta herramienta se ha encontrado un hiperplano de separación óptima con muy pocos clientes en las fronteras. Con la base de datos completa se obtienen 15 vectores en la frontera aceptados y 26 vectores en la frontera, y eliminado la variable Días de atraso de la última cuota pagada de la base de datos, se han obtenido 14 vectores en la frontera de aceptados y 14 vectores en la frontera de rechazados. En ambos casos es reducido el número de vectores soporte, por lo que se puede concluir que el criterio de selección usado por la microfinanciera está dando lugar a una separación de 2 grupos de clientes bien definidos.

(10) El uso de estas tres máquinas de aprendizaje ayuda a mejorar la atención a la cartera de clientes, porque en la BD estudiada, la micro-financiera toma la decisión solo en función de la variable V8. Si eliminamos esta variable, se debería escoger la variable mejor correlacionada y hacerla intervenir en la toma de decisiones.

RECOMENDACIONES

(1) La predicción del comportamiento crediticio de los clientes se recomienda hacerlo con Redes Neuronales Artificiales Backpropagation de 4 capas con 14, 10, 8 y 1 neuronas respectivamente. Con esta arquitectura de RNA se ha obtenido la mayor precisión en las pruebas realizadas.

(2) Para agrupar los clientes en dos clústeres, se recomienda usar una red SOM de dos neuronas ya sea con topología Gridtop o Hextop y métrica Dist o Linkdist. Con cualquiera de estas redes SOM se obtiene una alta rentabilidad y un riesgo bajo.

(3) Para agrupar los clientes en tres clústeres, se recomienda usar una red SOM de tres neuronas con topología Hextop y métrica Linkdist. Con esta red SOM se obtiene la rentabilidad más alta en el clúster C1 que contiene al grupo más numeroso de clientes. El riesgo con esta red SOM también es bajo.

(4) El uso de estas tecnologías obliga a que la micro-financiera haga una reingeniería en su BD, en la que participen en la decisión otras variables que no se están tomando en cuenta.

(5) El uso de las máquinas de aprendizaje exige la mejora permanente de la BD. No se puede seguir trabajando con una BD tal cual, en la que las decisiones se toman en función de una sola variable.

REFERENCIAS BIBLIOGRÁFICAS

Abid, L., Masmoudi, A., & Zouari-Ghorbel, S. (2016). The Consumer Loan's Payment Default Predictive Model: an Application of the Logistic Regression and the Discriminant Analysis in a Tunisian Commercial Bank. *Journal of the Knowledge Economy*, 9(3), 948-962. Doi:10.1007/s13132-016-0382-8

Ali AghaeiRad, A., Chen, N., & Ribeiro, B. (2016). Improve credit scoring using transfer of learned knowledge from self-organizing map. *Neural Computing and Applications*, 28(6), 1329-1342. DOI: 10.1007/s00521-016-2567-2

Ala'raj, M., & Abbod, M. F. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach. *Expert Systems With Applications*, 64, 36-55. DOI: 10.1016/j.eswa.2016.07.017

Ala'raj, M., & Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, 104, 89-105. DOI: 10.1016/j.knosys.2016.04.013

Armaki, A., Fallah, M., Alborzi, M., & Mohammadzadeh, A. (2017). A Hybrid Meta-Learner Technique for Credit Scoring of Banks' Customers. *Engineering, Technology & Applied Science Research* Vol. 7, No. 5, 2017, 2073-2082.

Asbanc (2019). Estadísticas del Sistema Financiero. Recuperado de https://www.asbanc.com.pe/Paginas/Estadistica/Estadisticas.aspx?gclid=EAIaIQobChMIntSvxd_s5QIVDYiGCh1WDAAzEAAYASAAEgKDv_D_BwE

Baklouti, I. (2013). Determinants of Microcredit Repayment: The Case of Tunisian Microfinance Bank. *African Development Review*, 25(3), 370–382.

Barboza, G., & Trejos, S. (2009). Micro Credit in Chiapas, México: Poverty Reduction Through Group Lending. *Journal of Business Ethics*, 88(S2), 283-299.

Bekhet, H. A., & Eletter, S. F. K. (2012). Credit Risk Management for the Jordanian Commercial Banks: A business Intelligence Approach. *Australian Journal of Basic and Applied Sciences*, 6(9): 188-195, 2012.

Bekhet, H. A., & Eletter, S. F. K. (2014). Credit risk assessment model for Jordanian commercial banks: Neural scoring approach. *Review of Development Finance*, 4(2014), 20-28.

Beltran Pascual, M., Muñoz Martinez, A., & Muñoz Alamillos, Á. (2014). Redes Bayesianas aplicadas a problemas de credit scoring. Una aplicación práctica. *Cuadernos de Economía*, 37(104), 73-86.

Bequé, A., & Lessmann, S. (2017). Extreme Learning Machines for Credit Scoring: An Empirical Evaluation. *Expert Systems with Applications*, 86, 42-53. doi:10.1016/j.eswa.2017.05.050

Blanco, A., Pino-Mejías, R., Lara, J., & Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from Peru. *Expert Systems with Applications,* 40(1), 356–364.

Carmona Suárez Enrique (2014). Tutorial sobre Máquinas de Vectores Soporte (SVM). http://www.ia.uned.es/~ejcarmona/publicaciones/[2013-Carmona]%20SVM.pdf

Chen, F.-L., & Li, F.-C. (2010). Combination of feature selection approaches with SVM in credit scoring. *Expert Systems with Applications*, 37 (7), 4902–4909.

Cubiles-De-La-Vega, M.-D., Blanco-Oliver, A., Pino-Mejías, R., & Lara-Rubio, J. (2013). Improving the management of microfinance institutions by using credit scoring models based on Statistical Learning techniques. *Expert Systems with Applications*, 40(17), 6910–6917.

Derelioglu, G., & Gürgen, F. (2011). Knowledge discovery using neural approach for SME's credit risk analysis problem in Turkey. *Expert Systems with Applications*, 38(8), 9313-9318.

Dixon, R., Ritchie, J., & Siwale, J. (2007). Loan officers and loan "delinquency" in Microfinance: A Zambian case. Accounting Forum, 31(1), 47-71.

Dželihodžić, A., Đonko, D., & Kevrić, J. (2018). Improved Credit Scoring Model Based on Bagging Neural Network. International Journal of Information Technology & Decision Making, 1–17. DOI:10.1142/S0219622018500293

Espinoza, P. (2020). Separata del Curso Aplicaciones Cuantitativas para la Investigación. Capítulo 8, Máquinas con Soporte Vectorial, 167-184.

Feng, X., Xiao, Z., Zhong, B., Qiu, J., & Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability. *Applied Soft Computing*, 65, 139-151.

Finlay, S. (2010). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, 210 (2011) 368–378.

Gulati, R., Goswami, A., & Kumar, S. (2018). What drives credit risk in the Indian banking industry? An empirical investigation. Economic Systems. doi:10.1016/j.ecosys.2018.08.004

Ghosh, S. (2018). Loan delinquency in banking systems: How effective are credit reporting systems? *Research in International Business and Finance*. 47(2019) 220-236 doi: 10.1016/j.ribaf.2018.07.011

Han, L., Han, L., & Zhao, H. (2013). Orthogonal support vector machine for credit scoring. Engineering Applications of Artificial Intelligence, 26(2013), 848–862.

Harris, T. (2015). Credit scoring using the clustered support vector machine. *Expert Systems with Applications*, 42(2015), 741–750.

Hillier, F., Lieberman, F. Introducción a la Investigación de Operaciones. Editorial Mc Graw Hill, 9na. Edición, México 2010.

Hsieh, N. (2005). Hybrid mining approach in the design of credit scoring models. *Expert Systems with Applications*, 28 (2005) 655–665.

Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. Expert Systems with Applications, 37(2010), 6233-6239.

Khashman, A. (2011). Credit risk evaluation using neural networks: Emotional versus conventional models. *Applied Soft Computing*, 11(2011), 5477-5484.

Koutanaei, F. N., Sajedi, H., & Khanbabaei M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*, 27(2015), 11-23.

Kiruthika, & Dilsha, M. (2015). A Neural Network Approach for Microfinance Credit Scoring. Journal of Statistics and Management Systems, Vol. 18 (2015), No. 1 & 2, pp. 121-138.

Leong, C. K. (2015). Credit Risk Scoring with Bayesian Network Models. *Computational Economics*, 47(3), 423-446.

Levenberg, K. (1944). A Method for the Solution of Certain Problems in Lest Squares. Quart. Appl. Math. Vol. 2, pp 164-168.

Liu, Z., & Pan, S. (2017). Fuzzy-Rough Instance Selection Combined with Effective Classifiers in Credit Scoring. Neural Processing Letters, 47(1), 193-202. doi: 10.1007/s11063-017-9641-3

Malhotra, R., & Malhotra, D.K. (2014). Identifying Potential Loan Defaulters in The Credit Union Environment: a Comparative Analysis Of Statistical and Neural Network Models. *Journal of Information Technology Case and Application Research*, 2(2), 20-48.

Martínez, M (2004) Yunus: Microcrédito, sobreendeudamiento y retos del sector. Portal de Microfinanzas. Recuperado de http://www.microfinancegateway.org/es/library/yunus-microcr%C3%A9dito-sobreendeudamiento-y-retos-del-sector

Melo Junior, L., Maria Nardini, F., Renso, C., Trani, R., & Antonio Macedo, J. (2020). A novel approach to define the local region of dynamic selection techniques in imbalanced credit scoring problems. *Expert Systems with Applications*, 113351. doi:10.1016/j.eswa.2020.113351

Moradi, S., & Mokhatab Rafiei, F. (2019). A dynamic credit risk assessment model with data mining techniques: evidence form Iranian banks. *Financial Innovation*, 5(1). doi:10.1186/s40854-019-0121-9

Nanayakkara, G., & Stewart, J. (2015). Gender and other repayment determinants of microfinancing in Indonesia and Sri Lanka. *International Journal of Social Economics*, 42(4), 322-339.

Nurlybayeva, K., & Balakayeva, G. (2013). Algorithmic Scoring Models. Applied Mathematical Science, Vol. 7, 2013, no. 12, 571-586.

Oreski, S. (2014). Hybrid Techniques of Combinatorial Optimization with Application to Retail Credit Risk Assessment. *ARTIFICIAL INTELLIGENCE AND APPLICATIONS*, 1(1), 21-43.

Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert Systems with Applications*, 41(4), 2052-2064.

Pai, P.-F., Tan, Y.-S., & Hsu, M.-F. (2015). Credit Rating Analysis by the Decision-Tree Support Vector Machine with Ensemble Strategies. *International Journal of Fuzzy Systems,* 17(4), 521-530.

Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems,* 118 (2019) 33–45.

Patricio Figueroa M. (2014). Teorías heurísticas que sustentan la capacidad de resolución de problemas. https://www.matematicas.cl/teorias-heuristicas/

Portocarrero, F., Trivelli, C., & Alvarado, J. (2002). Microcrédito en el Perú: quiénes piden, quiénes dan. Recuperado de https://www.cies.org.pe/sites/default/files/files/diagnosticoypropuesta/archivos/dyp-09.pdf

Rayo S., Lara J., & Camino D. (2010). Un Modelo de Credit Scoring para instituciones de microfinanzas en el marco de Basilea II. Journal of Economics, Finance and Administrative Science. Vol. 15 N° 28, 89-124.

Reyes Purizaca Jorge Luis (2017). La importancia de las Mypes sostenibles y su inserción en la Economía Peruana. https://es.slideshare.net/JorgeReyes103/importancia-de-las-mypes-sostenibles-y-su-insercin-en-la-economa-peruana

Shi, J., Zhang, S., & Qiu, L. (2013). Credit scoring by feature-weighted support vector machines. *Journal of Zhejiang University-SCIENCE C*, 14(3), 197-204.

Taha, H. Investigación de Operaciones. Pearson Education, 9na.Edición, México 2012.

Tian, Y., Yong, Z., & Luo, J. (2018). A new approach for reject inference in credit scoring using kernel-free fuzzy quadratic surface support vector machines. *Applied Soft Computing Journal*, 73, 96-105. https://doi.org/10.1016/j.asoc.2018.08.021

Tripathi, D., Edla, D. R., Cheruku, R., & Kuppili V. (2019). A novel hybrid credit scoring model based on ensemble feature selection and multilayer ensemble classification. *Computational Intelligence*. 2019;1–24. DOI: 10.1111/coin.12200

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert System with Applications*, 38(1), 223-230.

Zhang H., He H., & Zhang W. (2018). Classifier selection and clustering with fuzzy assignment in ensemble model for credit scoring. *Neurocomputing*, 316, 210-221. https://doi.org/10.1016/j.neucom.2018.07.070

Zhao, Z., Xu, S., Kang, B.H., Kabir, M. M. J., Liu, Y., & Wasinger, R. (2015). Investigation and improvement of multi-layer perception neural networks for credit scoring. *Expert Systems with Applications*, 42(7), 3508-3516.

Zhou, L., Lai, K.K., & Yu, L. (2008). Credit scoring using support vector machines with direct search for parameters selection. *Soft Comput* (2009) 13:149–155.

Zhou, X., Jiang, W., Shi, Y., & Tian, Y. (2011). Credit risk evaluation with kernel-based affine subspace nearest points learning method. *Expert Systems with Applications*, 38 (2011), 4272–4279

ANEXOS

ANEXO 1: DICCIONARIO DE DATOS

N°	Variable	Nombre	Tipo de Dato	Validación	Total de control	Consideraciones
1	**V1**	Tipo de moneda	Alfanumérico			Moneda de la operación 1: Soles, 2: Dólares
2	**V2**	Monto del crédito otorgado	Numérico			Monto equivalente en moneda nacional, con dos dígitos decimales. En caso de créditos revolventes colocar el monto de la línea de crédito aprobada.
3	**V3**	Saldo capital de la deuda	Numérico		Indicar suma total de todos los registros del campo	Monto neto equivalente en moneda nacional, con dos dígitos decimales. Equivalente a la suma de las colocaciones directas e indirectas y detrayendo los ingresos diferidos
4	**V4**	Tipo de crédito según Reporte Crediticio de Deudores	Numérico		Información debe coincidir con los saldos de las colocaciones reportadas en el RCD	01: Créditos Soberanos, 02: Créditos a Entidades del sector público, 03: Créditos a Bancos multilaterales de desarrollo, 04: Créditos a Empresas del sistema financiero, 05: Créditos a Empresas de valores, 06: Créditos Corporativos, 07: Créditos a Grandes Empresas, 08: Créditos a Medianas Empresas, 09: Créditos a Pequeñas Empresas, 10: Créditos a Microempresas, 11: Créditos de

						Consumo revolventes, 12: Créditos de Consumo no revolventes, 13: Créditos Hipotecarios para vivienda.
5	**V5**	Clasificación del deudor, según Anexo N° 5	Numérico	Información debe coincidir con Anexo N° 5	Información debe coincidir con el Anexo N° 5	0: Normal, 1: CPP, 2: Deficiente, 3: Dudoso, 4: Pérdida Considera el factor de la SBS (es un ponderado)
6	**V6**	Clasificación del deudor sin considerar alineamiento con el sistema	Numérico			0: Normal, 1: CPP, 2: Deficiente, 3: Dudoso, 4: Pérdida Es una clasificación propia, no considera la SBS
7	**V7**	Días de atraso al cierre de mes	Numérico			Calculados respecto a la fecha de corte y la fecha de vencimiento puntual.
8	**V8**	Días de atraso de la última cuota pagada	Numérico	Si no canceló cuotas o paga anticipadamente reportar 0.		Para todos los clientes que hayan cancelado al menos una cuota, reportar los días de atraso reales desde la última cuota pagada.
						Fecha de pago última cuota cancelada-fecha de vencimiento última cuota cancelada
9	**V9**	Promedio de días de atraso incurridos en el pago de cuotas correspondientes a los últimas 6 meses	Numérico			Para todas las cuotas con fecha de vencimiento durante los últimos 6 meses, desde la fecha de corte, considerar:
						Días de atraso por cuota/(Nº de cuotas)
						Solo considerar para créditos no revolventes

10	**V10**	Provisión constituida	Numérico	Debe coincidir con el Anexo N° 5	Indicar suma total de todos los registros del campo	Monto equivalente en moneda nacional, con dos dígitos decimales. Considerar provisiones genéricas obligatorias (considerar pro cíclica, si corresponde) y específicas, de acuerdo a la Norma SBS N°11356-2008.
11	**V11**	Saldo de capital vigente de la operación	Numérico	El total debe ser igual a la cuenta 1401	Indicar suma total de todos los registros del campo	Monto equivalente en moneda nacional, con dos dígitos decimales
12	**V12**	Saldo de capital vencido de la operación	Numérico	El total debe ser igual a la cuenta 1405	Indicar suma total de todos los registros del campo	Monto equivalente en moneda nacional, con dos dígitos decimales (es el nro. de cuotas capital no pagado)
13	**V13**	Saldo de capital en cobranza judicial de la operación	Numérico	El total debe ser igual a la cuenta 1406	Indicar suma total de todos los registros del campo	Monto equivalente en moneda nacional, con dos dígitos decimales
14	**V14**	Rendimientos devengados de la operación	Numérico	El total debe ser igual al saldo 1408	Indicar suma total de todos los registros del campo	Monto equivalente en moneda nacional, con dos dígitos decimales
15	**V15**	Intereses en suspenso acumulados de la operación	Numérico	El total debe ser igual al saldo de cuenta de orden respectiva (810401, 02, etc.)	Indicar suma total de todos los registros del campo	Monto equivalente en moneda nacional, con dos dígitos decimales, si el crédito es vencido, refinanciado o judicial
16	**V16**	Fecha de desembolso	Fecha			En formato DD/MM/AAAA
17	**V17**		Alfanumérico			1: Pago único de Capital e Intereses

		Esquema de amortización				2: Pago único de Capital pero con pago intermedio de Intereses. 3: Pago en Cuotas Fijas (Capital más Intereses
						4: Pago en Cuotas con capital constante 5: Otros
18	**V18**	Número de días de gracia para pago de capital según cronograma	Numérico			"Considerar como días de gracia el número de periodos de gracia multiplicados por la periodicidad de las cuotas"
19	**V19**	Fecha de vencimiento general de la operación	Fecha			En formato DD/MM/AAAA Fecha en la cual se cancela la totalidad del capital e intereses de la operación, según el cronograma pactado.
20	**V20**	Fecha de vencimiento puntual de la operación	Fecha			En formato DD/MM/AAAA - Operaciones con pago único: fecha de vencimiento de la operación.
						- Operaciones con pago en cuotas: fecha de vencimiento de la próxima cuota a vencer; salvo existan cuotas vencidas, en cuyo caso se deberá reportar la fecha de vencimiento de la cuota más antigua no pagada.
21	**V21**	Periodicidad de cuotas	Numérico			Indicar el número de días entre cuotas. En caso no aplique por el esquema de amortización consignar cero.
22	**V22**	Número de cuotas programadas	Numérico			De acuerdo al cronograma inicial. Consignar 0 en caso no aplique.

23	**V23**	Número de cuotas pagadas	Numérico			Cuotas pagadas hasta la fecha de corte. Consignar 0 en caso no aplique.
24	**V24**	Indicador de RFA	Numérico			Si deudor está refinanciado bajo el Rescate Financiero Agropecuario: 0 Si RFA; 1: No RFA
25	**V25**	Código de agencia	Alfanumérico			Código interno de la agencia u oficina a cargo de la operación. Adjuntar leyenda.
26	**V26**	Tasa efectiva anual	Numérico			Tasa de interés efectivo anual
27	**V27**	Flag de rechazo	Numérico			0: Se acepta, 1: Se rechaza