UNS
ESCUELA DE
POSTGRADO

EFICACIA DE LOS MODELOS DE APRENDIZAJE DE MAQUINA PARA EVALUAR EL RIESGO CREDITICIO DE PERSONAS NATURALES EN UNA INSTITUCIÓN FINANCIERA DE CHICLAYO

TESIS PARA OBTENER EL GRADO ACADÉMICO DE DOCTOR EN ESTADÍSTICA MATEMÁTICA

AUTOR:

M. Sc. Alfonso Tesén Arroyo

ASESOR:

Dr. Javier Arturo Gamboa Cruzado

CHIMBOTE - PERU
2017

UNS
UNIVERSIDAD
NACIONAL DEL SANTA

CONSTANCIA DE ASESORAMIENTO DE LA TESIS DOCTORAL

Yo, **JAVIER ARTUROGAMBOA RUZADO**, mediante la presente certifico mi asesoramiento de la Tesis Doctoral titulada: **EFICACIA DE LOS MODELOS DE APRENDIZAJE DE MAQUINA PARA EVALUAR EL RIESGO CREDITICIO DE PERSONAS NATURALES EN UNA INSTITUCION FINANCIERA DE CHICLAYO**, elaborada por el Magister ALFONSO TESEN ARROYO para obtener el Grado Académico de Doctor en ESTADÍSTICA MATEMÁTICA en la Escuela de Postgrado de la Universidad Nacional del Santa.

Nuevo Chimbote, 25 de mayo del 2017.

.....

DR. JAVIER ARTURO GAMBOA CRUZADO
ASESOR

UNS
UNIVERSIDAD
NACIONAL DEL SANTA

HOJA DE CONFORMIDAD DEL JURADO EVALUADOR

EFICACIA DE LOS MODELOS DE APRENDIZAJE DE MAQUINA PARA EVALUAR EL RIESGO CREDITICIO DE PERSONAS NATURALES EN UNA INSTITUCION FINANCIERA DE CHICLAYO

TESIS PARA OPTAR EL GRADO DE DOCTOR EN ESTADÍSTICA MATEMÁTICA

Revisado y Aprobado por el Jurado Evaluador

.....
DR. CARLOS ALBERTO MINCHON MEDINA
PRESIDENTE

.....
DR LUIS A. RUBIO JACOBO
SECRETARIO

.....
DR. JAVIER A. GAMBOA CRUZADO
VOCAL

Agradecimiento

Agradezco a Dios por estar conmigo en cada paso que doy

A mis padres y hermanos de los cuales siempre recibí su apoyo

A mi asesor de Tesis Dr. Erwin Kraenau Espinanal. (QEPD) con quien empezamos este trabajo juntos y desde el cielo vio su culminación.

A mi asesor Dr. Javier Gamboa Cruzado por estar siempre en la disposición de ofrecerme su ayuda, conocimientos, aportes y tiempo en este trabajo

Gracias a todo aquel que de una manera u otra intervino para que esta tesis hoy sea una realidad.

ÍNDICE.

LISTADO DE CUADROS

LISTADO DE GRÁFICOS

RESUMEN

La presente investigación tuvo como objetivo determinar que los modelos de aprendizaje de máquina evalúan eficazmente el riesgo crediticio de personas naturales de una institución financiera de Chiclayo que el modelo clásico de credit scoring estimado mediante la Regresión Logística.

La investigación es de tipo descriptivo, explicativo y predictivo, para lo cual se trabajó con la metodología CRISP- DM.

Para el desarrollo de la investigación se utilizaron los modelos de aprendizaje de maquina tales como, Árboles de Clasificación, Redes Neuronales, Maquinas de Soporte Vectorial y el modelo clásico de la Regresión Logística; la base de datos estuvo constituida por 2464 clientes, de los cuales se utilizó el 70% de la base para el entrenamiento y el 30% restante para la validación. Para la comparación de los modelos, se utilizó la Matriz de Confusión y la curva ROC, determinando que el mejor modelo de clasificación global en la etapa de entrenamiento fue la Redes Neuronales con un 81.10% y 82% en la etapa de validación; mientras que el mejor modelo de estimación del riesgo crediticio se obtuvo mediante el árbol de decisión para nuestros datos planteados con un 35.30% y 32,21% en las etapas de entrenamiento y validación respectivamente.

Finalmente se concluyó que los modelos de aprendizaje de maquina evalúan mejor el riesgo crediticio que el modelo de enfoque paramétrico de la Regresión Logística para nuestros datos financieros.

Palabra clave: *Máquinas de Aprendizaje, Machine Learning*, Regresión Logístico, Redes Neuronales Artificiales, Arboles de decisión.

Llegando a la conclusión qué el mejor modelo planteado que nos otorga la mejor estimación y pronóstico de los morosos es el árbol de para nuestros datos planteados.

En la siguiente tabla comparativa se muestra las diferentes áreas bajo la curva de los modelos planteados, la cual como se verifico las redes neuronales obtuvo el mejor porcentaje de acierto y la cual se demuestra en el área bajo la curva que obtuvo la mejor el mayor porcentaje.

Cuadro N°16: Tabla comparativa 3

	Tabla comparativa del área bajo la curva					
	Logística	**Redes neuronales**	**Árbol de decisión**	**SVM-radial**	**SVM-lineal**	**polinomial**
Técnica	0.894	0.898	0.892	0.872	0.891	0.873

Fuente: Elaboración Propia

Con lo que se concluye que los modelos de aprendizaje de maquina entre ellos el de redes neuronales obtuvo mejores resultados que el de utilizar un enfoque paramétrico, como es el de regresión logística para nuestros datos financieros.

ABSTRACT

The objective of this research was to determine that machine learning models assess the credit risk of natural persons in a financial institution of Chiclayo more effectively than the classic model of credit scoring estimated using logistic regression.

This research is descriptive, explicative and predictive, for which we worked with the CRISP-DM methodology.

We used machine learning models such as decision trees, neural networks, support vector machines, and the classical logistic regression model; the database consisted of 2464 clients, with 70% of them used for training and 30% for validation. To compare the models, confusion matrices and ROC curves were used, and we determined that the best global classification model in the training stage was the neural network with a 81.10% and 82% in the validation stage; whereas the best model for credit risk estimation for our database were decision trees with 35.30% and 32.21% in the training and validation stages, respectively.

Finally, we conclude that, for our financial data, machine learning methods assess credit risk better than logistic regression parametric approach models.

Keywords: Machine learning, Logistic regression, artificial neural networks, decision trees.

CAPÍTULO I: INTRODUCCIÓN

1.1. REALIDAD PROBLEMÁTICA

El problema que enfrentan casi todas las entidades financieras, es la existencia de un nivel de riesgo en el cual estos entes están inmersos en la presencia de morosidad por parte de los prestatarios y hasta cierto punto de incobrabilidad de las operaciones al crédito resulta cada vez más preocupante por la incertidumbre que genera. (Marzo, C.; Wicijowski, C. & Rodriguez, L. 2008)

La gestión corporativa del riesgo se ha convertido en un elemento importante dentro de las políticas administrativas de las instituciones dedicadas al otorgamiento de créditos; de igual manera en el entorno económico peruano han cobrado especial importancia las figuras crediticias, como instrumentos para la generación de alternativas de crecimiento, teniendo claro que el crédito por sí solo no es suficiente para impulsar el desarrollo económico. (Aguilar, G. 2004)

Por diversas cuestiones existe un grupo identificable de agentes que la tecnología convencional en la producción de servicios financieros rechaza, principalmente por problemas de información; es en este sentido que adquiere relevancia la incertidumbre sobre el reembolso de los préstamos, cuando esta es elevada es posible que simplemente los préstamos no sean otorgados. Por otro lado, tradicionalmente se presume demasiado costoso adquirir la información que se necesita para pronosticar mejor el comportamiento del deudor. Siendo el interés del individuo no estar sujeto a dicha restricción, tiene incentivos para procurar proveer esta información al prestamista, estos problemas clasificados por la teoría como de información imperfecta, son causales de racionamiento del crédito. Mientras exista requerimiento de garantías colaterales y el costo de la información de selección y del monitoreo sea alto, los prestamistas convencionales restringirán los montos a disposición de los prestatarios o directamente podrían decidir no prestarle a determinado grupo de solicitantes. (Baca, G. A. 1997)

Así podemos identificar grupos de la población que debido a su bajo nivel de ingresos presentes y/o flujo futuro de ingresos inciertos, es decir, poseen escasos

activos liquidables, encuentran restringido su acceso al crédito a cualquier tasa de interés y esto implica un límite al nivel de bienestar alcanzable. La tecnología de crédito convencional implica un alto costo financiero para suplir la demanda de algunos tipos de créditos. Es así como cada vez más, los mercados de capitales se han vuelto accesibles a las Pequeñas y Medianas Empresas PYMES, y éstas a su vez no presentan estructuras financieras que faciliten su inserción en modelos de análisis de riesgo, presentándose inicialmente como altamente riesgosas por sí mismas, por ello, a través de una adecuada determinación de perfiles, pueden verse como un mercado potencial interesante que genere un buen índice de rentabilidad al atenderlas financieramente. (Krugman, R. P. 1995)

Vistas estas problemáticas se entiende que la restricción al acceso al crédito no necesariamente reflejará falta de capacidad de pago del potencial deudor, sino un complejo entramado de relaciones entre los prestamistas y los aspirantes a crédito; para contrarrestar los efectos adversos de la operatoria tradicional sobre la distribución del ingreso, se vuelve necesario explorar y desarrollar nuevas tecnologías de crédito que superen estas barreras. (Fragoso, J. 2002)

Se ha logrado encontrar la manera de solucionar el problema técnico de producir servicios financieros para clientelas marginadas a un costo razonable y una tasa de ganancia positiva, es decir, una función de producción (tecnología) que ha posibilitado este resultado, abriendo nuevas posibilidades para relajar las restricciones de liquidez. (Baca, G. A. 1997)

Dada la necesidad de establecer mecanismos para medir la probabilidad de no pago, se deben determinar las variables significativas que expliquen el fenómeno y contribuyan a generar un modelo de medición de riesgo de crédito, a través de métodos estadísticos teórico-prácticos, que puedan ser implementados para el otorgamiento de créditos en una en la Cooperativa de Ahorro y Crédito, haciéndose ampliamente pertinente dado que la institución está interesada en generar modelos con los cuales logre medir la posibilidad de no pago, en las diferentes líneas de créditos establecidas. Bajo las situaciones anteriores, se presenta la oportunidad de generar un método de medición de riesgo de crédito, el cual se abordará desde

tres modelos que permitan estimar la probabilidad de incumplimiento, con los cuales se puedan efectuar comparaciones de las bondades y desventajas que cada uno de ellos presenta. Para este efecto, la Entidad suministrará la información acerca del perfil de los clientes actuales en cada línea de crédito que permita asumir variables de tipo tanto cualitativo como cuantitativo y así desarrollar los modelos; bajo un laborioso proceso metodológico para el diseño y selección de una muestra, la cual generará los valores para las variables exógenas y endógenas con las cuales se correrá cada uno de los modelos. (Cabrera, A. 2014)

1.2. ESTADO DEL ARTE DEL TEMA DE LA INVESTIGACIÓN

La problemática que enfrenta casi todas las entidades financieras es la existencia de un nivel de riesgo en el cual estos entes están inmersos, la presencia de morosidad por parte de los prestatarios y hasta cierto punto de incobrabilidad de las operaciones al crédito que realizan son motivos por los cuales la colocación del crédito resulta cada vez más preocupante por la incertidumbre que genera.

La cooperativa de ahorro y crédito no es ajena a esta problemática ya que presenta dificultades en cuando a la concesión de créditos personales; actualmente sus clientes presentan retrasos con las fechas de pago, en algunos casos se enfrentan al incumplimiento total en el pago, situación en el cual debe optar por una serie de medidas que van de las vías administrativas (notificaciones y cobranzas a domicilio) a las vías judiciales (procesos de embargo) para poder evitar el continuo sobreendeudamiento. En estos últimos años la cooperativa presentó un nivel de morosidad muy variable, el año 2009 alcanzó un 8,21% que disminuyo hasta 7,94% al cierre del 2010, para fines del 2011 presento una morosidad de 6,76% y se incrementó a 9,75% al cierre del 2012 y que fue disminuyendo hasta un 4,79% para Junio del 2013. Esta variabilidad, demuestra una ineficiencia en la asignación de créditos y la necesidad de ajustar los criterios de evaluación de los clientes a fin de mantener una morosidad moderada en los próximos años.

El problema inicia en si cuando el crédito otorgado no es recuperado, por lo que la decisión de conceder o no el crédito se toma difícil al no contar con una herramienta que permita identificar a los clientes "buenos" y "malos" estimando así, la probabilidad de que un cliente incumpla con el crédito personal.

Cabrera, A. (2014) en su investigación titulada "Diseño de creditscoring para evaluar el riesgo crediticio en una entidad de ahorro y crédito popular", sostiene que: En caso de la institución objeto de estudio, se identificó una base de datos que no se encuentra depurada y contiene información que no aportó nada a la investigación. Las variables que se incluyeron en el modelo son propias de la institución y probablemente no sean útiles para otro modelo de creditscoring, puesto que lo que es significativo para una entidad no lo es para otra. El creditscoring representa para la institución una herramienta útil en la evaluación del sujeto a manera de sugerencias de aceptación o no de la solicitud de crédito. El nivel de riesgo que la institución esté dispuesto a correr será el indicador para aceptar o rechazar solicitudes, por otro lado, la determinación de los puntos de corte depende de cada institución. Se concluye que la hipótesis planteada se cumple al comprobarse que el perfil del cliente tiene que ver con su nivel de cumplimiento, obteniéndose que las variables que mejor pueden explicar la morosidad de la institución, de acuerdo al diseño del modelo de creditscoring obtenido son: Oficina, Buró, Producto y Estado civil. Por lo anterior, podemos determinar que dentro del creditscoring el perfil del cliente tiene que ver con su nivel de cumplimiento, puesto que logra diferenciar entre un cliente bueno y un cliente malo.

Ladino, I. (Marzo 2014) en su investigación titulada: "Comparación de modelos de riesgo de crédito: modelos logísticos y redes neuronales", sostiene que: Como resultado de comparar el desempeño de los modelos de ANN y la regresión logística en 1000 muestras con remplazo (bootstrap), los estadísticos D y C de los modelos ANN son estadísticamente mayores. Adicionalmente, la significancia económica en magnitud de esta mayor discriminación (0,0088 y 0,0052 en los estadísticos D y c respectivamente) es material, al disminuir la perdida en 4,4% para el caso analizado.

Es importante resaltar que es más difícil implementar los modelos ANN que los modelos logísticos en los sistemas de las entidades financieras. Los modelos ANN tienen un mayor poder de discriminación y un impacto económico material, sin embargo, este resultado es producto de comparar estos modelos con una función de costos simétrica. Un posible trabajo futuro consistiría en comparar los modelos logísticos y neuronales utilizando una función de costos asimétrica que podría otorgar mayor costo al error de clasificar un cliente malo como bueno, respecto al error de clasificar un cliente bueno como malo.

Moreno, S. (2013) en su investigación titulada: "El modelo Logit Mixto para la construcción de un scoring de crédito", sostiene que: Los tres modelos estimados (Logit tradicional, probit y logit Mixto) tiene un buen poder discriminatorio, reflejado en las altas tasas de aciertos sobre todo para los clientes morosos. El modelo logit mixto resultó ser el de mayor sensibilidad, aunque también predijo el mayor número de falsos positivos. En cuanto a las variables que determinan que un cliente llegue a default, resultaron significativas las relacionadas con el factor de comportamiento crediticio, financiero y demográfico, como se esperaba. Se descartaron algunas variables como las moras mayores a 30 días, por problemas de tipificación. Las variables que explican el evento de llegar a default en los modelos ajustados, resultaron con signos acordes con la realidad de la entidad financiera. Mediante el modelo logit mixto se pudo determinar que los factores nivel de estudio, tipo de relación laboral, número de meses desde el último crédito y edad definida por grupos, tiene un efecto aleatorio en el modelo para predecir el default en una entidad financiera del sector cooperativo. Para una entidad financiera es muy importante contar con una herramienta estadística adecuada para la predicción del comportamiento de los clientes al momento de otorgarles el crédito, puesto que la rentabilidad y los flujos de caja, en gran medida corresponden al correcto pago de las obligaciones. Crediticias contraídas por parte de los clientes. El modelo logit mixto es la más potente en la predicción o detección de los clientes que llegan al estado de default, pero esta predicción está asociada a que es un modelo muy estricto de la aceptación de clientes óptimos (no default), lo que genera un gran

Mallo, F. (2011) en su investigación titulada: "Modelos multivariantes internos de medición de riesgo de crédito acordes con Basilea II", sostiene que: En este trabajo proponemos una alternativa llamada Modelos Logísticos Lineales Híbridos de Expansiones Lineales por funciones de base (modelos HLLM). Los cuales se obtienen al expandir la componente no lineal de modelos logísticos parcialmente lineales, LPLM, a través de expansiones lineales en funciones de base específicas para cada variable. El modelo HLLM obtenido es parsimonioso y tanto en la base de entrenamiento, como en la validación y test, presenta características adecuadas no solo desde el punto de vista del riesgo de crédito, sino también desde el punto de vista estadístico: alta bondad de ajuste, un alto poder discriminante, alta eficacia como clasificador y un bajo error test apoderado. Además, presenta mejor rendimiento discriminante, que las técnicas utilizadas usualmente en los sistemas de calificación del riesgo de crédito, TREE, SLPM, K-NN Y SVM. Los modelos HLLM Y HLPM presentan menores tasas de clasificación incorrecta. Con respecto a la calibración, ninguno de los modelos está calibrado adecuadamente, si bien se tiene para todos ellos que la categoría de default y la de menor puntuación entre las no default no están infra estimadas.

Cabrera Jiménez, Juan Manuel y Pérez Pérez, Fabricio O. (México, 2010), en su proyecto de investigación titulada: "Clasificación de Documentos usando Naive Bayes Multinomial y Representaciones Distribucionales".

Se observa las ventajas de la representación distribucional DOR frente a otras formas de representación en tareas de clasificación de documentos, se compara las ventajas de los Clasificadores Bayesianos Multinomiales frente a los bayesianos simples. Primero, se filtra con la bolsa de palabras y se utilizan dos Clasificadores Bayesianos (NB y NBM).

La primera parte del procesamiento incluye dividir los datos incorpora Web KB en dos grupos, uno de entrenamiento y uno de pruebas, la razón es que los pesos de las representaciones utilizadas (tf=idf) emplean información de la colección, por lo

que si utilizamos toda la colección tendríamos información adicional referente al conjunto de pruebas, situación que alteraría los resultados. Después de dividir los datos en dos subconjuntos, el siguiente paso es realizar la representación de los documentos, es decir, se crean matrices tanto para la representación de bolsas de palabras como para la representación distribucional DOR. En dicha investigación, se utiliza Matlab versión 2010, además de Text to Matrix Generator (TMG).

Echeverri Valdés, Fanery (2010), cuya investigación es: "Evaluación de modelos para la medición de riesgo de incumplimiento en créditos para una entidad financiera del eje cafetero". A través del Modelo de Probabilidad Lineal, se logra observar las características principales de una relación no evidente entre las variables, determinando la probabilidad de ocurrencia o no, y del pago oportuno de un crédito.

El modelo aplicado(MPL) determinó como variables explicativas se comportan en los meses vigentes de la obligación, el número de pagos al año, los días máximos de vencimiento y el porcentaje de incumplimiento en los pagos, habiéndose planteado la hipótesis de la importancia de esas variables para establecer variaciones en la probabilidad de incumplimiento.

Otro modelo utilizado es el Logit, el cual plantea la clasificación de los datos de acuerdo a características comunes. Genera un ordenamiento de la cartera y la estimación de probabilidades permitiendo calcular las respectivas provisiones, se solucionan los problemas del modelo lineal a través del uso de funciones de distribución.

El modelo Discriminante analizó las diferencias entre grupos de datos y explico las variables que logran discriminar, es decir presentar la heterogeneidad de los grupos, las variables significantes fueron: los meses vigentes de obligación, los días máximos de vencimiento y el porcentaje de incumplimiento en los pagos. Sin embargo, este modelo tiene limitantes para tratar de calcular la probabilidad de incumplimiento.

Joekes, Silvia (Colombia, 2010) en su investigación titulada: "Modelado y pronóstico de una serie de tiempo contaminada empleando redes neuronales y procedimientos estadísticos tradicionales" emplearon las redes neuronales en la serie contaminada de RESEX utilizando el método de la Red Neuronal Artificial de backpropagation y comparándolos con los procedimientos estadísticos tradicionales.

Nieto, S. (Mayo 2010) en su investigación titulada: "Crédito al consumo: La estadística aplicada a un problema de riesgo crediticio", sostiene que: El primer paso en el proceso de CreditScoring es la depuración y preparación de la base de datos. La limpieza de la base de datos es un proceso largo y tedioso que se aligera con el uso de software adecuado para ello, en este caso nos auxiliamos de EXCEL y ACCESS de Microsoft; de acuerdo a los resultados obtenidos al estimar la matriz de transición, se decidió que los clientes buenos son aquellos que al final de los seis meses estaban al día en sus pagos y como máximo estuvieron dos pagos vencidos durante el periodo de seis meses. También son bueno clientes los que al final de los seis meses tienen un pago vencido y durante los seis meses tuvieron como máximo un pago vencido, este resultado es acorde con lo que se esperaba, pues los buenos clientes deben ser los que pagan y no entran en mora muchas veces. Los malos clientes son los que tienes tres o más pagos vencidos al final de los seis meses. Se puede utilizar la propiedad de las matrices de transición para discriminar a los buenos clientes de los malos clientes. Si se obtiene las potencias de la matriz de transición, se puede estimar la probabilidad que de cualquier estado se caiga en cartera vencida en dos, tres o más pasos, la decisión de cuáles pueden ser buenos o malos clientes se haría de manera semejante a lo realizado en esta tesis, esto también queda para trabajos futuros.

Vigo, G. (2010) en su investigación titulada: "Método de clasificación para evaluar el riesgo crediticio: Una comparación", sostiene que: No existe diferencias en el porcentaje de error de entretenimiento en los métodos: regresión logística y árbol

de clasificación (CART), utilizados para la clasificación de los clientes que solicitan un préstamo; sin embargo, con las redes neuronales se obtuvo un 84,18% de buena clasificación y un 74,32% de buena predicción. Con el modelo de regresión logística, se obtuvo mayor error en la clasificación y predicción debido a que este método es sensible a los valores influyentes, al igual que los árboles de clasificación (CART), por otro lado, la red neuronal es insensible a valores influyentes. Para trabajos posteriores se recomiendo utilizar meta modelos, que es la combinación de modelos diferentes, que subsanan el problema del sobreajuste de los datos de entretenimiento y que pueden mejorar la clasificación y la predicción.

Herrán, L. (julio 2009) en su investigación titulada: "Evaluación crediticia aplicando un modelo de CreditScoring en el ámbito microempresarial: caso CMAC PAITA", sostiene que: Respecto al CreditScoring, su elaboración debe estar basada, en principios, en variables relativas tanto al negocio como a las características personales. El Scoring obtenido en base al modelo Logit, permitió obtener resultados bastante aceptables (R2 de Mc Fadden de 0,52 y 0,78 de predicciones correctas). El Scoring aún es mejorable a partir de la constatación de que aún existen algunas debilidades en la información de esta entidad financiera. En particular, se careció de algunos datos para determinar la rentabilidad real al negocio financiero, ya que los flujos de caja son proyecciones elaboradas al momento de la aprobación de la solicitud de crédito. La variable más significativa en el modelo Logit final es NOREF; altamente significativa (1%); la cual a su vez presenta el mayor efecto marginal promedio de la probabilidad de incumplimiento de pago. El modelo de Scoring permitió probar que de la cual concentración de la cartera de créditos Microempresariales (MES) de la CMAC Paita el 16,75% son no puntuales y el 6,17% de las predicciones son incorrectas. En concreto el trabajo se orientó a estimar la probabilidad de incumplimiento de pago de un cliente en función a una serie de características, utilizando la metodología del CreditScoring; la cual se emplea mayormente para evaluar individuos, pequeñas y medianas empresas; ya que las grandes se analizan con sistemas de rating. Una buena aproximación de

estas probabilidades resulta muy importante para que la CMAC Paita reduzca sus pérdidas de morosidad.

Gustavo Maradona (2007), en dicha investigación cuyo título es: "Resultados de una aproximación preliminar a la predicción de series financieras utilizando redes neuronales". Se muestra una red neuronal con el fin de predecir el índice Merval en Argentina y comparar su desempeño con el de un camino aleatorio. En este trabajo utilizamos estos modelos de redes neuronales artificiales con el objetivo de predecir una serie financiera. En particular, la serie financiera predicha es el índice bursátil de Argentina: el Merval y los insumos utilizados para predecirla son: la tasa de interés, el valor del dólar, la inflación, el tipo de cambio real, el día de la semana, la actividad productiva industrial y el índice bursátil de Estados Unidos. Utiliza una red neuronal de tipo recurrente (Total-Recurrent) con dos capas ocultas. El algoritmo de aprendizaje es el de optimización de Powells. En los nodos de la primera capa oculta se utiliza la función de activación arco-tangente, lineal umbral en los de la segunda capa oculta y sigmoidea en la capa de output. Se utilizan como insumos, dos rezagos de la tasa de interés pasiva que reporta el BCRA y un rezago del Merval. Se concluye que las redes estimadas no logran un desempeño superior al obtenido en la aplicación de un modelo de camino aleatorio, presentando evidencia a favor de la hipótesis de mercados eficientes.

Gutiérrez Girault, Matías Alfredo (Argentina, 2007), dicha investigación titulada: "Modelos de CreditScoring: Qué, Cómo, Cuándo y Para qué". Se plantea un modelo, en la cual la variable dependiente toma valores discretos, se emplean modelos de regresión discreta. El caso más simple se da cuando ella es binaria y toma los valores 0 a 1, y se puede estimar con distintos enfoques como el modelo de probabilidad lineal, análisis discriminante, los modelos de tipo probit y logit o con una regresión logística. Se utilizó los datos del CENDEU, se construye un modelo que predice el comportamiento de los deudores retail del sistema financiero: individuos y PyMEs.

Resendiz Trejo, Juan Ángel (México, 2006), en su investigación titulada: "Las máquinas de vectores de soporte para identificar en línea". Se propone un método SVM recursivo para identificación en línea el cual admita en su representación a un conjunto de datos de entrenamiento los cuales no puedan ser representados por una combinación lineal de los datos de entrenamiento. El método propuesto es capaz de realizar de manera recursiva la identificación en líneas de sistemas no lineales.

Bensic, M., Sarlija, & N. Zekic - Susac, M. (2005) en su investigación titulada: "Modelización de la puntuación de crédito de la pequeña empresa mediante el uso de regresión logística, redes neuronales y árboles de decisión", sostiene que: El trabajo tuvo como objetivo comparar el rendimiento de las metodologías de las redes neuronales (RNA), regresión logística (LR) y los árboles de decisión (CART), así como para identificar características importantes para el modelo de calificación de crédito de la pequeña empresa en un conjunto de datos croata. Las medidas de asociación estadística mostrando que el mejor modelo de redes neuronales (RNA) está mejor relacionado con los datos que los modelos de RL y CART. El mejor modelo RN superó significativamente al modelo LR y extrajo características personales y de la empresa, así como características del programa de crédito como características importantes. Dado que el algoritmo probabilístico NN está diseñado específicamente para los problemas de clasificación de acuerdo con una función de probabilidad, reconocido y recomendado como un clasificador eficiente en algunas otras áreas de aplicación, el hecho de superar a los otros modelos indica que este algoritmo podría ser una propuesta para los problemas de este tipo de calificación de crédito.

Cardona Hernández, Paola Andrea (Colombia, 2004), en su investigación titulada: "Aplicación de árboles de decisión en modelos de riesgo crediticio", define dos tipos de modelos para predecir la probabilidad de incumplimiento de pago, los cuales son: modelo de otorgamiento, en este modelo de iniciación, contiene 6 nodos terminales, en los cuales permite identificar 6 perfiles de riesgo, mientras que el

segundo tipo de modelo es el comportamiento, con el que se controla la maduración del crédito. Para esta investigación se utiliza los arboles de decisión binarios, que son métodos no paramétricos. Los árboles de decisión se presentan como herramienta efectiva para la predicción de incumplimiento, y esto ayuda a una mejora en la provisión de la entidad financiera, y el grado de morosidad del cliente.

1.3. CARACTERIZACIÓN Y NATURALEZA DEL OBJETO DE INVESTIGACIÓN

De alcance explicativo y predictivo de corte transversal de enfoque cuantitativo.

1.4. FORMULACIÓN DEL PROBLEMA

¿Evalúan mejor los modelos de aprendizaje de máquina, el riesgo crediticio de personas naturales de una institución financiera de Chiclayo que el modelo clásico de creditscoring estimado mediante la Regresión Logística?

1.5. FORMULACIÓN DE LA HIPÓTESIS

Los modelos de aprendizaje de máquina sí evalúan mejor el riesgo crediticio de personas naturales de una institución financiera de Chiclayo que el modelo clásico de creditscoring clásico estimado mediante la Regresión Logística.

1.6. FORMULACIÓN DE LOS OBJETIVOS DE LA INVESTIGACIÓN

1.6.1. OBJETIVO GENERAL

Determinar que los modelos de aprendizaje de máquina evalúan más eficazmente el riesgo crediticio de personas naturales de una institución financiera de Chiclayo que el modelo clásico de creditscoring estimado mediante la Regresión Logística.

1.6.2. OBJETIVO ESPECÍFICOS

Estimar el riesgo crediticio mediante los modelos de aprendizaje de máquina: redes neuronales, máquinas de soporte vectorial, árboles de clasificación y el modelo clásico de creditscoring estimado mediante la Regresión Logística.

Determinar que los modelos de aprendizaje de máquinas evalúan mejor el riesgo crediticio de personas naturales de una institución financiera de Chiclayo que el modelo clásico de creditscoring estimado mediante la Regresión Logística.

1.7. IMPORTANCIA Y JUSTIFICACIÓN DE LA INVESTIGACIÓN

Hoy en día, estimar y administración el riesgo financiero estaba en manos de los analistas de crédito quienes cuantificaban el riesgo solamente a través de un conjunto de reglas propias de cada institución. El explosivo aumento de las solicitudes crediticias, la incorporación de nuevos requerimientos y el almacenamiento de información en grandes bases de datos no permitían a los analistas realizar procesos manuales rápidos y eficientes para la concesión de créditos; esto sumando a la creciente tasa de morosidad, obligaron a diversas instituciones financieras a implementar diferentes medidas con el fin de poder evaluar las pérdidas esperadas frente al incumplimiento de pago por parte del cliente. La medida más utilizada hoy en día es el scoring de créditos o creditscoring.

El Banco Interamericano de Desarrollo (2010) señala que el creditscoring destaca como una herramienta de gran potencial para contribuir al control del riesgo y la utilización eficiente de recursos en dichas entidades, mejorando así su rendimiento financiero y social, así como la calidad de su cartera.

La cooperativa de ahorro y crédito objeto de estudio no es ajena a esta problemática, a lo largo de estos últimos años ha venido presentando preocupantes índices de morosidad y dificultades en cuanto al otorgamiento de créditos personales. La decisión de aprobar un crédito se toma difícil al no contar con una

herramienta que pueda predecir la probabilidad de incumplimiento de pago y que clasifique a sus clientes; por tales motivos es indispensable la construcción de un modelo de creditscoring para créditos personales.

La presente investigación se justifica desde el punto de vista teórico-práctico. Presenta justificación teórica porque generará reflexión y discusión sobre el conocimiento existente del área investigada, ya que por medio de las técnicas de regresión logística, árboles de clasificación y redes neuronales se estimaron modelos que permitén predecir la probabilidad de incumplimiento de pago; al mismo tiempo se confrontarán las técnicas trabajadas y se contrastarán los modelos obtenidos (en base a indicadores de eficiencia y predicción) con el objetivo de determinar el mejor modelo de creditscoring para el otorgamiento de crédito personal.
Presenta justificación práctica porque el modelo construido evaluará al cliente mediante un score y lo clasificará como un buen o mal pagador para un futuro crédito personal, contribuyendo de esta manera con la cooperativa de ahorro y crédito a realizar un adecuado proceso de otorgamiento de crédito y disminuir el índice de morosidad que presenta.

La importancia de la presente investigación radica en que un modelo de creditscoring adecuado favorecerá no solamente a la cooperativa de ahorro y crédito sino también a los clientes en sí. Entre los beneficios se destaca el poder mejorar la calidad de la cartera crediticia y el servicio de crédito personal que brinda; el ayudar al analista de crédito en la toma de decisiones de una forma más rápida y objetiva basándose en probabilidades y no solo por el juicio humano al momento de otorgar un crédito, la prevención de pérdidas futuras y la predicción del comportamiento de futuros prestamos con el fin de reducir los índices de morosidad.

En cuanto a su alcance, esta investigación servirá como marco de referencia para la realización de posteriores estudios en la región de Lambayeque sobre el riesgo de crédito en entidades financieras. Por otra parte, sirve como base para que los

estudiantes de la escuela profesional de estadística de la UNPRG utilicen las diversas técnicas estadísticas para desarrollar investigaciones en campos actuales y poco explorados como en Big Data, Data Mining, Business Intelligence, entre otros.

CAPÍTULO II: MARCO TEÓRICO

2.1. RIESGO CREDITICIO

La principal actividad de una entidad financiera es aquella que mejor la define y a la que dedica la mayor parte de sus esfuerzos. La actividad que genera la mayor parte de sus beneficios y los mayores riesgos, es la actividad crediticia. (Vigo, G. 2010)
Habitualmente la palabra riesgo tiene una connotación negativa: algo que debemos evitar. Sin embargo, el negocio bancario supone precisamente eso, la gestión de riesgos con el objetivo de obtener una rentabilidad que compense adecuadamente. Un banco es básicamente una máquina de gestión de riesgos en busca de rentabilidad. De todos los riesgos a los que está expuesto el negocio bancario, el principal es el riesgo de crédito. Este se define como la posibilidad de incurrir en pérdidas como consecuencia del incumplimiento por parte del deudor de sus obligaciones en las operaciones de intermediación crediticia. El más grave de los incumplimientos es el impago. (Vigo, G. 2010)

2.2. ADMINISTARCION DEL RIESGO CREDITICIO

La Entidad Financiera debe contar al menos con los componentes básicos del Sistema de Administración del Riesgo Crediticio (SARC): políticas claras de administración de riesgos, una estructura organizacional adecuada, metodologías y procesos apropiados para la gestión de riesgos, así como un proceso de auditoría general. (Baca, G. A. 1997)
Políticas y Estructura Organizacional: La definición de una política clara de administración de riesgo por parte de la Junta Directiva de la entidad, constituye el eje central del SARC. Esta política debe reflejar el nivel de tolerancia frente al riesgo dado al nivel de rentabilidad esperado, generando límites para las distintas exposiciones del portafolio de crédito, acordes con el capital de respaldo. Asimismo, la Junta debe garantizar o exigir a la administración de la entidad la asignación adecuada de tiempo y recursos físicos y humanos para el cumplimiento de esta política, así como reportes sobre los niveles de exposición, las implicaciones de los mismos y las actividades relevantes para su mitigación y/o gestión.(Baca,G.A. 1997)

2.3. EL SISTEMA FINANCIERO NACIONAL

El Sistema Financiero cumple un rol fundamental en el desarrollo y crecimiento de una economía. A través de la intermediación de fondos, los Sistemas Financieros eficientes generan asignaciones de activos óptimas entre los agentes de la economía, lo cual permite expandir la frontera de producción y alcanzar mayores niveles de utilidad, es decir, mejorar el nivel del bienestar social. (Levi, D. M. 1997)

Constituye el marco institucional que pone en contacto a los agentes económicos (familias, empresas, estado) ofertantes y demandantes de fondos prestables (ahorro), para efectuar transacciones financieras de captación y aplicación de fondos. Tradicionalmente el Sistema o Mercado Financiero se puede clasificar en dos: por un lado el Sistema de Intermediación Indirecta (SII), Sistema Bancario o Mercado de Dinero cuya institución de supervisión y control es la Superintendencia de Banca y Seguros; y por el otro lado el Sistema de Intermediación Directa (SID), Sistema No Bancario o Mercado de Valores cuya institución de supervisión y control es SMV (Superintendencia del Mercado de Valores). (Galicia, M. 2003)

2.4. ANÁLISIS DE CRÉDITO

El análisis de crédito es considerado como un arte; ya que no hay esquemas rígidos y que por el contrario es dinámico y exige creatividad, Sin embargo resulta importante dominar las diferentes técnicas de análisis de créditos y complementarla con una amplia experiencia y buen criterio, asimismo es necesario contar información disponible, necesaria y suficiente que permita minimizar el número de incógnitas para poder tomar la decisión correcta. (Nieto, S. 2010)

2.5. TÉCNICAS DE CLASIFICACIÓN

La mayoría de las técnicas de minería de datos (TMD) existentes son para la tarea de clasificación. Estas técnicas se usan para el aprendizaje supervisado; lo cual implica, que las observaciones del conjunto de entrenamiento están previamente

agrupadas por una variable denominada "variable clase" que se modela en función de un conjunto de variables predictoras. Las técnicas de clasificación pueden resolver muchos problemas en diferentes campos: medicina, industria, negocio, educación, ciencias, etc. La clasificación es el proceso de encontrar un modelo o función que describa datos etiquetados con alguna clase, con el propósito de ser utilizado para predecir datos nuevos de una clase que es desconocida (S. Venkata Krishna Kumar & P. Kiruthika, 2015).

2.5.1. Métodos para la evaluación de las técnicas de clasificación

Para medir la performance de las técnicas de clasificación se han propuesto una serie de métodos y criterios con la finalidad de validar y evaluar su bondad de ajuste al conjunto de datos, cuya aplicación dependerá de la TMD empleada.

Entre los métodos propuestos de tienen:

2.5.1.1. Métodos para la validación de clasificadores

Son métodos que permiten evaluar la performance de los clasificadores, cuya finalidad es realizar una evaluación honesta sobre su bondad de ajuste al conjunto de datos. Los métodos consisten en dividir el conjunto total de observaciones en tres subconjuntos: conjunto de entrenamiento (usado para el proceso de aprendizaje o estimación del clasificador), conjunto de validación (usado para validar el clasificador) y conjunto de prueba (usado para la inferencia de nuevas observaciones); el ultimo conjunto es opcional y generalmente se usa datos no incluidos en la base de datos. La regla de clasificación se encuentra al determinar un punto de corte entre 0 y 1. Si para un elemento la probabilidad de que Y=1 es mayor que el punto de corte; se considera que este se ubica en la clase de interés A, determinada por Y=1; de otro modo el elemento se ubica en la clase determinada por Y=0. (Hernández, O. R. 2015)

Existen varios métodos para validación de los clasificadores tales como: el método de validación cruzada, la tabla de confusión, la curva ROC, entre otras.

2.5.1.1.1. Métodos de validación cruzada (Croos – Validation)

Es el más utilizado en el aprendizaje supervisado, consiste en dividir aleatoriamente el conjunto de entrenamiento D en k subconjuntos (k-folds) mutuamente excluyentes $\{D_1, D_2, \ldots, D_k\}$ de similar tamaño. El proceso de validación cruzada es repetido durante k iteraciones, de tal manera que en cada iteración el clasificador usa un subconjunto para la validación (D_V) y es entrenado con los k – 1 subconjuntos $(D - D_V)$, el error de clasificación se calcula como la media aritmética de los errores de cada iteración. Un caso particular de la validación cruzada dejar – uno – afuera (Leave – one–out), implica que en cada iteración se tenga un solo dato de prueba y el resto para entrenamiento, el error se calcula como el promedio de los errores cometidos. (Peña, D. 2002)

2.5.1.1.2. Holt – Out

Este método particiona aleatoriamente el conjunto de datos D en dos conjuntos mutuamente excluyentes: conjunto de entrenamiento (D_E) y conjunto de validación (D_V). El tamaño de D_E generalmente es mayor al D_V en proporciones 2/3 y 1/3, 4/5 y 1/5, etc. respectivamente. Los elementos del D_E suelen obtenerse mediante muestreo sin reemplazo de todo el conjunto de datos, mientras que el conjunto D_V lo conforma las observaciones restantes que no pertenecen al D_E. Suele ser aplicado a un conjunto de datos. (Perez, M. 2014).

2.5.1.1.3. Tabla de confusión

La tabla de confusión se usa para evaluar y comparar la confiabilidad de las técnicas de clasificación. Se basa en estimar los porcentajes de la correcta e incorrecta clasificación (la tasa de aciertos y error) que

realiza el clasificador con el conjunto de datos. La tabla de confusión, es una tabla de contingencia que muestra la distribución del número de observaciones que están correcta e incorrecta clasificadas, con respecto a la clasificación observada y la predicha por el clasificador, considerando las distintas categorías de la variable clase. (Véliz, C. 2016)

Cuadro N°1: Tabla de confusión

Clasificación observada (Valores de Y)	Clasificación predecida (Valores pronosticados de Y)		Total (Observado)
	Positiva (+1)	Negativa (-1)	
Positiva (+1)	VP	FN	VP+FN
Negativa (-1)	FP	VN	FP+VN
Total (Predecido)	VP+FP	FN+VN	VP+FN+FP+VN

Para interpretar mejor los resultados que aparecen en la tabla, se supone que los valores de la variable dependiente Y son: +1 cuando sucede el evento de interés A y -1 cuando esto no ocurre (esta nomenclatura se usa generalmente en epidemiología para modelar los resultados de pruebas de laboratorio y para las cuales el evento de interés es el resultado positivo).

Las frecuencias en la tabla indican lo siguiente:

VP= Número de predicciones correctas para los valores +1
Cada predicción así realizada se llama verdadera positiva.

FN= Número de predicciones incorrectas para los valores +1.
Cada predicción así realizada se llama falsa negativa

FP= Número de predicciones incorrectas para los valores -1.
Cada predicción así realizada se llama falsa positiva.

VN= Número de predicciones correctas para los valores -1
Cada predicción así realizada se llama verdadera negativa.

Al utilizar estas frecuencias se obtiene las siguientes medidas:

La precisión o capacidad de acierto total del modelo que se expresa como:

$$CAT = \frac{VP + VN}{VP + FN + FP + VN}$$

La tasa verdadera positiva o sensitividad del modelo que se expresa como:

$$v_p = \frac{VP}{VP + FN}$$

La tasa falsa positiva del modelo que se expresa como:

$$f_p = \frac{FP}{FP + VN}$$

La tasa verdadera negativa o especificidad del modelo que se expresa como:

$$v_n = \frac{VN}{FP + VN}$$

La tasa falsa negativa del modelo que se expresa como:

$$f_n = \frac{FN}{VP + FN}$$

2.5.1.1.4. Análisis de las curvas ROC

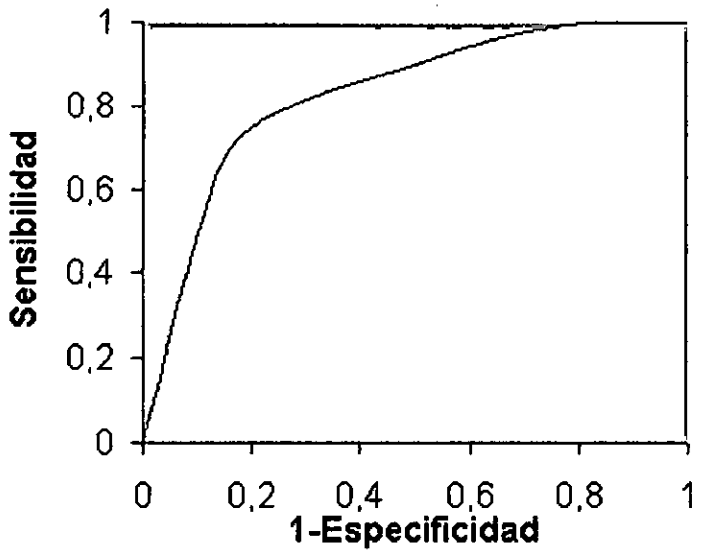
El análisis de la curva ROC (Receiver Operating Characteristic), es una curva que fue introducida para estudiar la detección de señales. La curva ROC es útil para comparar el comportamiento global de un modelo y solo es posible su representación con variables respuesta binaria. Este gráfico enfrenta dos variables: la sensibilidad (sensitivity) y 1 = especificidad (1 = specificity). (Véliz, C. 2016)

- **Sensitivity:** es una medida de la capacidad de acierto de un evento y se define como el número de categorías positivas (categoría de referencia) bien predichas dividido por el total de categorías positivas.

- **Specificity:** es una medida de la capacidad de acierto del evento complementario al anterior. Se define como el número de categorías falsas (categoría complementaria a la referencia) bien predichas (el modelo también dice que son falsas) dividido por el total de categorías falsas. 1 – Specificity es simplemente el número de falsos positivos o número de observaciones con categoría falsa que el modelo incorrectamente predice como verdaderas para un punto de corte determinado dividido por el número de casos falsos.

La curva ROC permite determinar un punto de corte de tal modo que los valores de la probabilidad mayores que el indica que Y=1.
En la siguiente figura se representa la curva ROC correspondiente a diferentes puntos de corte de un modelo.

Grafico N°1: Curva ROC



Si el modelo es perfecto, hay una región en la que cualquier punto de corte tiene sensibilidad y especificidad iguales a 1 y la curva tiene solo el punto (0, 1).

Si el modelo no ayuda en la clasificación, la sensitividad es igual a la tasa de falsos positivos (1 - especificidad) y la curva es la diagonal que va de (0, 0) a (1, 1).

El modelo tiene mejor desempeño si la curva ROC correspondiente se aleja más de la diagonal principal.

Cuadro N°2: Interpretación del área bajo la curva ROC

	Poder Discriminante
Área ROC =0.5	Nulo (como lanzar una moneda)
0.7 ≤ Área ROC < 0.8	Aceptable
0.8 ≤ Área ROC < 0.9	Excelente
Área ROC ≥ 0.9	Excepcionalmente buena

Autor: Carlos Véliz Capuñay

2.5.1.1.5. Capacidad predictiva de un modelo

Para evaluar la capacidad predictiva de un modelo se utilizan varios estadísticos alternativos. (Pérez, M. 2014) Siendo n el horizonte de predicción, los estadísticos más habituales para la evaluación de la capacidad predictiva son los siguientes:

Raíz del error cuadrático medio (Rott Mean Squared Error):

$$RECM = \sqrt{\frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{n}}$$

Error Absoluto medio (Mean Absolute Error):

$$EAM = \frac{\sum_{i=1°}^{n} |\hat{Y}_i - Y_i|}{n}$$

Error absoluto medio del porcentaje del error (Mean Abs. Percent Error):

$$EAMP = \frac{1}{n} \sum_{i=1°}^{n} \left| \frac{\hat{Y}_i - Y_i}{Y_i} \right|$$

Coeficiente de desigualdad de Theil (TheilInequalityCoefficient):

$$CDT = \frac{\sqrt{\frac{\sum_{i=1°}^{n} (\hat{Y}_i - Y_i)^2}{n}}}{\sqrt{\frac{\sum_{i=1°}^{n} \hat{Y}_i^2}{n}} + \sqrt{\frac{\sum_{i=1°}^{n} Y_i^2}{n}}}$$

Proporción del sesgo (BiasProportion):

$$\frac{(\bar{\hat{Y}} - Y)^2}{\sum_{i=1°}^{n} (\hat{Y}_i - Y_i)^2 / n}$$

Proporción de la varianza (VarianceProportion):

$$\frac{(S_{\hat{Y}} - S_Y)^2}{\sum_{i=1°}^{n} (\hat{Y}_i - Y_i)^2 / n}$$

Proporción de la covarianza (CovarianceProportion):

$$\frac{2(1 - r) S_{\hat{Y}} S_Y}{\sum_{i=1°}^{n} (\hat{Y}_i - Y_i)^2 / n}$$

Cuanto más próximos estén a cero los valores de los cuatro primeros estadísticos, mejor será la capacidad predictiva del modelo, lo que

permitirá comparar un modelo con otros alternativos. Las tres proporciones varían entre cero y uno y también es conveniente que sean pequeñas.

2.5.1.2. Medidas para la selección de modelos

Si se tiene un conjunto de modelos M1, M2, ... con parámetros K1, K2, ... respectivamente, dos medidas para compararlos son el Criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC). Ambos criterios usan el Log-Verosimilitud y penalizan la complejidad del modelo.

2.5.1.2.1. El criterio de información de Akaike (AIC)

Akaike propuso un enfoque alternativo para resolver el problema de seleccionar el modelo suponiendo que el objetivo es hacer predicciones tan precisas como sea posible. Sea $f(y|M_i)$ la densidad de una nueva observación bajo el modelo M_i y sea $f(y)$ la verdadera función de densidad, que puede o no ser uno de los M_i. Para seleccionar el modelo de manera que $f(y|M_i)$ sea tan próxima como sea posible a $f(y)$. (Peña, D. 2002) Una manera razonable de medir la distancia entre estas dos funciones de densidad es mediante la divergencia de Kullback – Leibler entre las dos densidades, que se calcula mediante:

$$KL(f(y|M_i), f(y)) = \int log\frac{f(y|M_i)}{f(y)} f(y) d_y$$

Para interpretar esta medida observemos que la diferencia de logaritmo equivale a la diferencia relativa, ya que

$$log\frac{f(y|M_i)}{f(y)} = \log\left(1 + \frac{f(y|M_i) - f(y)}{f(y)}\right) \cong \frac{f(y|M_i) - f(y)}{f(y)}$$

Y cuando las diferencias son grandes, el logaritmo es mejor medida de discrepancia que la diferencia relativa. Una manera alternativa de escribir esta medida es:

$$KL\big(f(y|M_i), f(y)\big) = E_y \log f(y|M_i) - E_y \log f(y)$$

Donde E_y indica obtener la esperanza bajo la verdadera distribución de y. Como esta cantidad es siempre positiva, minimizaremos la distancia entre la verdadera distribución y $f(y|M_i)$ haciendo el primer término lo mas pequeño posible. Puede demostrarse, (Akaike, 1985), que esto equivale a minimizar

$$AIC = -2L(M_i) + 2p_i = D(M_i) + 2p_i$$

Donde p_i es el número de parámetros del modelo M_i. El criterio es minimizar la suma de la desviación del modelo, que disminuirá si introducimos más parámetros, más el doble del número de parámetros en el modelo, que tiende a corregir este efecto.

2.5.1.2.2. Criterio de BIC

El criterio de BIC (Bayesian Information Criterion) fue obtenida por primera vez por Schwarz (1978), que propuso escoger el modelo que conduzca a un valor máximo de esta cantidad. (Peña, D. 2002) Una forma equivalente de este criterio, es calcular para cada modelo la cantidad:

$$BIC\,(M_j) = -2L_j(\hat{\theta}_j|X) + p_j \log n$$

Y seleccionar aquel modelo para el cual esta cantidad es mínima. Este criterio pondera la desviación del modelo, medida por $-2L_j(\hat{\theta}_j|X)$, con el número de parámetros. Si se introducen más parámetros en el modelo mejorará el ajuste con lo que aumentará el soporte o disminuirá la

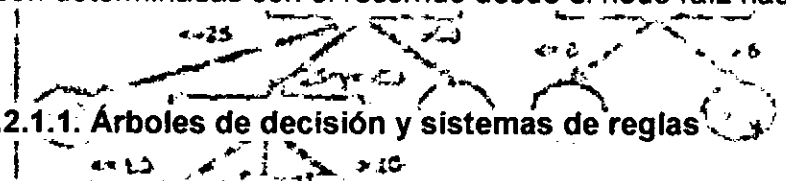
desviación, y este efecto queda compensado por el aumento del número
de parámetros que aparece en $p_j \log n$.

2.5.2. Tipos de técnicas de clasificación

2.5.2.1. Árboles de decisión

Un Árbol de clasificación (AC) es una técnica para el aprendizaje inductivo supervisado, fácil de implementar y a su vez de los más poderosos para problemas de clasificación. La estructura de un AC corresponde a un grafo acíclico dirigido, compuesto por un nodo llamado raíz, un conjunto de nodos internos que se les asocia una variable y cuyos arcos representan los diferentes valores que toma la variable que permiten la interconexión entre los nodos internos y los nodos terminales, llamados hojas de árbol que están etiquetadas con algún valor de la variable clase.

Los árboles de clasificación, son métodos más utilizados en el aprendizaje supervisado, por presentar una estructura jerárquica simple para comprender y tomar decisiones acerca de los datos (Lathi, R., Chitre, V., & Patil, N., 2012). El conocimiento se representa en el árbol a través de un conjunto de reglas (condiciones) organizadas en su estructura jerárquica, y que son determinadas con el recorrido desde el nodo raíz hasta las hojas.



2.5.2.1.1. Árboles de decisión y sistemas de reglas

De acuerdo a Hernándes, O. 2015, un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumple desde la raíz del árbol hasta alguna de sus hojas. Los árboles de decisión se utilizan desde hace siglos, y son especialmente apropiados para expresar procedimientos médicos, legales, comerciales, estratégicos, matemáticos, lógicos, etc.

de decisión en concreto funciona como un "clasificador", es decir, dado un nuevo individuo nos lo clasifica en una de las dos clases posibles: "no" o "sí".

Por otro lado, los sistemas de reglas son una generalización de los árboles de decisión en el que no se exige exclusión ni exhaustividad en las condiciones de las reglas (es decir, podría aplicarse más de una regla o ninguna).

La representación en forma de reglas suele ser, en general, más sucinta que la de los árboles, ya que permite englobar condiciones y permite el uso de reglas por defecto.

La diferencia más importante entre los sistemas de aprendizaje de árboles de decisión y los sistemas de inducción de reglas proposicionales es la filosofía del algoritmo que utilizan: participación o cobertura.

2.5.2.1.2. Estructura de árboles de decisión

Partiendo de una Base de Datos con una variable *Y* a discriminar, denominada variable respuesta, y un conjunto finito de variables X_1, X_2, ..., X_k conocidas como variables explicativas. Se tratará de seleccionar entre las variables explicativas aquellas que discriminen mejor a la variable *Y*. Obteniéndose una partición de la población de forma que se encuentren dos o más subgrupos lo más heterogéneos posibles entre sí con respecto a la variable respuesta *Y*, y lo más homogéneos posibles dentro. Esta discriminación se continúa para los nuevos nodos generados y se aplica un criterio de parada, obteniendo el árbol de clasificación o regresión. (Rodríguez, J. 2010) Todo árbol de clasificación comienza con un nodo al que pertenecen todos los casos de la muestra a clasificar (*nodo raíz*), el resto de nodos se dividen en *nodos intermedios* o no terminales y *nodos*

hojas o nodos terminales. Un árbol de decisión consta de los siguientes elementos:

Nodos intermedios: se generan dos o más segmentos descendientes inmediatos (dependiendo del método empleado). También llamados segmentos intermedios.

Nodos terminales: Es un nodo que no se puede dividir más. También denominado segmento terminal.

Rama de un nodo t: Consta de todos los segmentos descendientes de t, excluyendo t.

Árbol de decisión completo (Amax): Árbol en el cual cada nodo terminal no se puede ramificar.

Sub-árbol: Se obtiene de la poda de una o más ramas del árbol *Amax*.

A pesar de los distintos tipos de árboles de clasificación y regresión existentes la forma de actuar en todos ellos es similar, salvo ligeras modificaciones. En primer lugar se debe tener un conjunto de datos con una variable respuesta (categórica o continua) y un conjunto de variables explicativas, todas ellas categóricas o continuas que han sido previamente categorizadas. Todos los registros de la base de datos son examinados para encontrar la mejor regla de clasificación de la variable respuesta. Estas reglas se realizan basándose en los valores de las variables explicativas. La secuencia de particiones define el árbol. Cada partición se realiza para optimizar la clasificación del subconjunto de datos. El proceso de división es recursivo y finaliza la ramificación cuando se verifica un criterio de parada que ha debido ser definido previamente.

2.5.2.1.3. Árboles de decisión para regresión, agrupamiento o estimación de probabilidades

Aunque los árboles de decisión parecen idóneos para clasificación, se han adaptado para otras tareas, como son la regresión, el agrupamiento o la estimación de probabilidades. De hecho, uno de los primeros algoritmos de aprendizaje de árboles de decisión, el CART (Breiman et al. 1984), es tanto un clasificador como un árbol de regresión.

En realidad, un árbol de regresión se construye de manera similar a un árbol de decisión para clasificación, pero con las siguientes diferencias:

- La función aprendida tiene dominio real y no discreto, como en los clasificadores.
- Los nodos hoja del árbol se etiquetan con valores reales, de tal manera que una cierta medida de calidad se maximice, por ejemplo la varianza de los ejemplos que caen en ese nodo respecto al valor asignado.

La implementación más sencilla de esta idea es el propio algoritmo CART, que hace particiones binarias sobre los atributos de igual manera que lo visto para los árboles de decisión diseñados para clasificación, pero que va asignando una media y una varianza a cada nodo, intentando seleccionar las particiones que reduzcan las varianzas de los nodos hijos. Un sistema similar es CHAID (Kass 1980) (derivado de AID y THAID (Morgan &Sonquist 1963; Morgan & Messenger 1973)), que, en realidad, realiza particiones no binarias y usa test-*f* para determinar la partición óptima en el caso de regresión y test ji-cuadrado para clasificación.

Una variación muy popular de los árboles de regresión es considerar una función lineal en los nodos en vez de una media y una desviación típica. En este caso, en primer lugar, para evaluar las particiones se puede utilizar, por ejemplo, el error cuadrático medio de la regresión lineal de los ejemplos que hayan caído en cada nodo. En segundo lugar, para los nodos hoja, la predicción se realiza utilizando el modelo lineal. Nótese que

un modelo lineal se puede obtener de una forma relativamente sencilla y eficiente para un conjunto de ejemplos. También hay que destacar que esta modificación es directa sí todos los atributos son numéricos. En caso que existan también atributos nominales, se debería utilizar algún tipo de regresión lineal que trate con estos atributos nominales. (Rodríguez, J. 2010)

Además de su uso para regresión, los árboles de decisión han sido modificados para utilizarse en agrupamiento. La primera idea es modificar el criterio de partición y de evaluación para que considere particiones que separen entre zonas densas y poco densas. Esto se sigue haciendo hasta que se llega a zonas muy densas o zonas muy poco densas, constituyendo entonces los nodos del árbol. Los grupos formados corresponden a los nodos de las zonas densas. Un refinamiento de esta idea es el método presentado por (Liu et al. 2000), en el que se consideran todos los ejemplos de una clase "E" (por existentes) mientras que se añaden ejemplos ficticios "N" (por no existentes) uniformemente distribuidos en el espacio. El siguiente paso es simple, se utiliza un método de aprendizaje de árboles de decisión para clasificación (preferiblemente con poda). El resultado es que las reglas obtenidas para la clase "E" serán los *clusters* o grupos formados.

Evidentemente esta idea se puede refinar más y no es necesario crear realmente los puntos "N", sino que se puede rediseñar el algoritmo, en particular, el criterio de partición, para que los considere, como si estuvieran allí. No obstante, sino disponemos de un algoritmo de agrupamiento mediante árboles de decisión, podríamos generar los ejemplos ficticios nosotros mismos y utilizar algoritmos de clasificación clásicos, tales como CART o C4.5 para poder hacer el agrupamiento.

Finalmente, los árboles de decisión se pueden utilizar también para la estimación de probabilidades. En este caso, la presentación del problema es similar a la de un problema de clasificación; los ejemplos tienen una

etiqueta discreta denominada clase. La diferencia es que el objetivo de los estimadores de probabilidades es más ambicioso. No se trata de determinar para cada nuevo ejemplo de qué clase es, sino determinar para cada nuevo ejemplo cuál es la probabilidad de que pertenezca a cada una de las clases. Un estimador de probabilidades es especialmente útil cuando se quieren acompañar las predicciones con cierta fiabilidad, o se quieren combinar predicciones de varios clasificadores, o bien se quiere hacer un *ranking* de predicciones. (Hernández, O. 2008)

La modificación de un clasificador árbol de decisión para que sea un estimador de probabilidades es bien sencilla. Supongamos que tenemos tres clases *a, b* y *c*. Para cada nodo hoja con una cardinalidad *n* (número total de ejemplos de entrenamiento que caen en ese nodo) tendremos un determinado número de ejemplos de cada clase: n_a, n_b y n_c. Si dividimos cada uno de estos valores por la cardinalidad total, tendremos una estimación de las probabilidades de las clases en ese nodo, es decir; $p_a=n_a/n$, $p_b=n_b/n$ y $p_c=n_c/n$. Este tipo de árboles de decisión modificados de esta manera se denominan PETs (Probability Estimation Trees).

Aunque la conversión es sencilla, las estimaciones de probabilidad obtenidas por los PETs de esta manera no son muy buenas comparadas con otros métodos. No obstante,

- El suavizado de frecuencias de las estimaciones de probabilidad de las hojas, como por ejemplo la corrección de Laplace o el *m*-estimado, mejoran significativamente las estimaciones, especialmente si se utilizan para hacer *rankings*.
- La poda (o técnicas relacionadas de transformación o de colapsamiento) no es beneficiosa para la estimación de probabilidades. Los árboles sin podar dan los mejores resultados.

Teniendo en cuenta estas consideraciones, los árboles de decisión constituyen también una buena herramienta para la estimación de probabilidades.

Finalmente, los árboles de decisión se han utilizado frecuentemente como métodos de envolvente (*wrapper*) para la selección de atributos. En realidad, los criterios de partición eligen el atributo que sea más discriminante, es decir, más significativo. Utilizando los valores obtenidos por cualquiera de los criterios de partición podemos determinar un orden de significatividad de los atributos. Esto se puede realizar considerando sólo la partición al nivel superior del árbol (en la raíz) o estudiando el uso que se hace de los atributos en todas las particiones del árbol, ponderándolas convenientemente. Por ejemplo, el algoritmo CHAID, comentado anteriormente, se ha utilizado más para tareas de selección de atributos y detección de interacciones [Kass 1975] entre los atributos que para tareas de clasificación o regresión. De hecho, el nombre de CHAID proviene de Chi-squared Automatic Interaction Detector. (Hernandes, O. R. 2015)

2.5.2.1.4. Construcción de árboles de decisión

Para la construcción de árboles de decisión se deben tener ciertas etapas, estas son:

1. ***Construir el árbol (Reglas de división)***
   - ✓ Al inicio todos los ejemplos de entrenamiento están a la raíz.
   - ✓ Los atributos deben ser categóricos (si son continuos ellos deben ser discretizados)
   - ✓ El árbol es construido recursivamente de arriba hacia abajo con una visión de divide y conquista.
   - ✓ Los ejemplos son particionados en forma recursiva basado en los atributos seleccionados

- ✓ Los atributos son seleccionados basado en una medida heurística o estadística (ganancia de información)
- ✓ La ganancia de información se calcula desde el nivel de entropía de los datos.

2. ***Detener la construcción (Reglas de parada):***
   - ✓ Todas las muestras para un nodo dado pertenecen a la misma clase.
   - ✓ No existe ningunos atributos restantes para ser particionados (el voto de la mayoría es empleada para clasificar la hoja).
   - ✓ No existe más ejemplos para la hoja.

3. ***Podar el árbol (Reglas de poda)***
   - ✓ Identificar y eliminar ramas que reflejen ruido o valores atípicos.

2.5.2.1.5. Algoritmos

Los árboles de decisión generalmente son construidos con la ayuda de un algoritmo, el cual divide los registros en grupos; la probabilidad del resultado es diferente en cada grupo atendiendo a los valores de las variables independientes. Existe una gran variedad de algoritmos de árboles de decisión:

CHAID: El CHAID (Chi=squared Automatc Interaction Detector), fue creado por (Gordan, 1980). Es un algoritmo estadístico rápido y multidireccional que explora rápida y eficientemente datos, y construye segmentos y perfiles en función de la variable de respuesta establecida. El CHAID, es fácil de interpretar, manejar y se puede utilizar para la clasificación y detección de la interacción entre las variables. Fue introducido por Gordan V Kass en 1980, CHID es una extensión de la ayuda (Automatic Interaction Detector) y procedimientos THAID

(Interaction Detector Theta automática). El criterio para particionar está basado en $\chi 2$ y para terminar el proceso se requiere definir de antemano un "*threshold*" (umbral). Funciona sobre el principal de las pruebas de significación corregida. Después de la detección de la interacción entre las variables se selecciona el mejor atributo para dividir el nodo que hizo un nodo hijo como una colección de valores homogéneos de atributo seleccionado. El método puede manejar valores que faltan. Esto no implica que cualquier método de poda.

En (Patel & Rana, K., 2014) se presenta el estudio de los árboles de clasificación ID3, C4.5 y C5.0 respecto a las características, retos, ventajas y desventajas. Se concluye que el rendimiento de los algoritmos en estudio depende en alto grado de la medida entropía, ganancia de información y las variables del conjunto de datos. Asi mismo, algunos algoritmos reducen el problema de réplicas, manejo de datos continuos y el sesgo con variables con múltiples valores. En (Katare & Athavale, V. A., 2011) se presenta una revisión y estudio de los algoritmos de árboles de clasificación ID3, C4.5 y C5.0 proporcionando las propiedades básicas que afectan su performance en su construcción (sobe ajuste, poda, tamaño del árbol y valores missing en las variables). Además, se muestra las características diferencias del C5.0 con respecto a sus precesores: responde al ruido y datos missing, reduce el error de poda, disminuye el problema de sobre ajuste). En (Patil, Lathi, R., &Chitre, V., 2012) se comparan los algoritmos C5.0 y CART para apoyar el proceso de toma de decisiones para recomendar la afiliación de tarjetas de crédito. Se aplica ambos algoritmos a los registros de 7000 clientes, dividiendo en un 70% para el entrenamiento y 30% para la validación. Las reglas obtenidas son utilizadas para clasificar la afiliación de nuevos clientes.

CHAID Exhaustivo: Es una modificación del CHAID que examina todas las posibles particiones de la variable predictora.

Árboles de Clasificación y Regresión (C&RT ó CART): El CART (Classification and regressiontree), fue diseñado por L. Brieman. Es un algoritmo para construir árboles binarios de clasificación y regresión (variable clase categórica o continua), usando el Índice de Gini (IG) como la medida de impureza para seleccionar los atributos con la mayor reducción en la impureza que será divido en subconjuntos homogéneos precisos. CART puede trabajar con atributos numéricos y categóricos y con datos missing. El índice de Gini como medida de impureza para la selección de atributos. Existe una versión similar llamada Ind CART distribuido por la NASA. El criterio para particionar es la impureza del nodo. Carro acepta datos con valores numéricos o categóricos y también se encarga de valores de atributos que faltan. Se utiliza el coste – complejidad y también generar árboles de regresión.

QUEST: Es un algoritmo estadístico que selecciona variables de manera no sesgada y construye árboles binarios precisos rápida y eficientemente.

ID3: El ID3 (Iterative Dichotomiser 3), es un algoritmo de árbol de decisión desarrollado por (Quilan J. R., 1986). El ID3, es el algoritmo mas simple y potente para construir un árbol, trabajando solo con atributos nominales. El ID3 utiliza la Ganancia de Información (GI), como la medida para generar los nodos (raíz, intermedio y hoja) del árbol, seleccionando aquel atributo que proporcione la mayor ganancia de información (menor entropía y menor incertidumbre del atributo). El ID3, utiliza la Entropía (menor valor, menor incertidumbre y mayor información proporciona el atributo para la clasificación) y la Ganancia de Información (ganancia por usar el atributo) como medidas de la bondad de los atributos.

C4.5: El algoritmo C4.5 fue propuesto por J. Ross Quinlan (1993) dentro de la comunidad de "Machine Learning", siendo una extensión del ID3. El

C4.5, utiliza como medida la Razón de Ganancia (RG) y criterio para ir seleccionando el atributo que dividirá el conjunto de entrenamiento que definirán los nodos del árbol. El C4.5 realiza la poda del árbol después de haberlo construido (post poda) posibilitando tener arboles más consistentes y evitando el problema de sobre ajuste. Puede trabajar con atributos nominales y continuos; pero los atributos continuos son convertidos en intervalos discretos (discretización automática) y puede construir árboles cuando existen datos faltantes (missing).

NewId o C5.0: El algoritmo C5.0 es una extensión del algoritmo C4.5, que es también la extensión del ID3. El C5.0, puede ser aplicado a grandes volúmenes de datos (Big Data). El C5.0, supera al C4.5, en cuanto a la velocidad, la memoria y la eficiencia. El algoritmo C5.0, usa la máxima ganancia de información para dividir el conjunto de entrenamiento. El modelo C5.0 puede dividir la muestra en base a la información de campo de ganancia mayor. El subconjunto de la muestra que se obtiene de la primera división se dividirá después. El proceso continuara hasta que el subconjunto de la muestra no se puede dividir y por lo general es de acuerdo a otro campo. Por último, examinar la división de nivel más bajo, se rechazarán los subconjuntos de la muestra que no tienen notable contribución al modelo C5.0 se maneja fácilmente el atributo de valor múltiple y falta un atributo de conjunto de datos.

Árboles Bayesianos: Es un algoritmo basado en la aplicación de métodos Bayesianos a árboles de decisión. Buntine (1992).

2.5.2.1.6. Modelo de árbol de clasificación: CART

El vector de variables predictoras es X, donde algunas de las variables Xi son cualitativas y otras son cuantitativas. Entonces, el conjunto Q de preguntas binarias en los nodos debe tener las siguientes características:

a. Cada división de los nodos depende del valor de una sola variable predictora.

b. Si la variable Xi es continua entonces Q incluye todas las preguntas de la forma { Es Xi ≤ c }, donde c es cualquier número real. Usualmente c es el punto medio entre dos valores consecutivos de un atributo.

c. Si la variable Xi es categórica tomando valores en { b_1, b_2,....., b_m } entonces Q incluye todas las preguntas de la forma { Xi Є A? } donde A es un subconjunto cualquiera de { b_1,b_2,.....b_m }. En total se pueden considerar 2^{m-1}-1.

Un árbol de decisión particiona el espacio de variables predictoras en un conjunto de hiper-rectángulos y en cada uno de ellos ajusta un modelo sencillo, generalmente una constante.

Es decir y = c, donde y es la variable de respuesta. La idea fundamental es que los nodos hijos sean más puros que los nodos padres. La partición de un nodo t del árbol T se hace de acuerdo a un criterio que es diseñado para producir nodos hijos que produzcan una suma de cuadrados de errores menor que separen mejor las clases que el del nodo padre en el caso de clasificación.

En árboles de clasificación sean la proporción de observaciones en el nodo s, y la proporción de observaciones en el nodo s que pertenecen a la clase g (g = 1,.., G), donde G es el número de clases.

El índice de la impureza del nodo t como donde es una función de impureza, la cual debe satisfacer ciertas propiedades. Entonces la regla para particionar el nodo t es formar el nodo hijo derecho t_R y el nodo hijo izquierdo t_L tal que la disminución de la impureza dada por: sea máxima. Para árboles de clasificación se pueden usar las siguientes medidas de impureza:

Error de clasificación:

$$Error(t) = 1 - \max_j [p(j|t)]$$

Donde: p(j|t) = Es la probabilidad de pertenecer a la clase jestando en el nodo t

Índice de Gini:

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

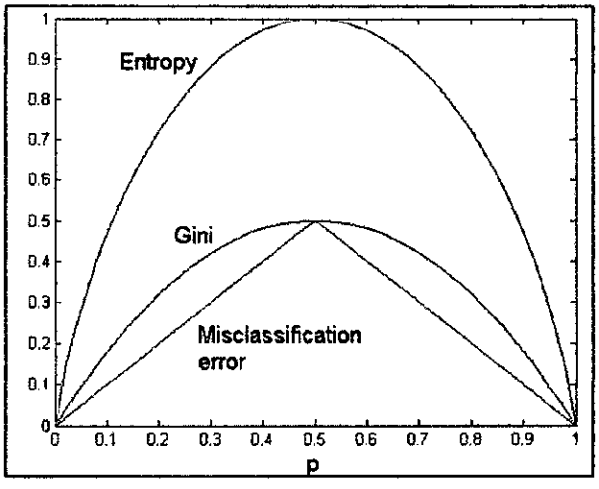
Donde:

P (t) es la proporción de caos en el nodo de t de la categoría i.

Entropía:

$$Entropía(t) = -\sum_j p(j|t) log_2 p(j|t)$$

Gráfico N°3: Comparación gráfica



2.5.2.2. Máquinas de soporte vectorial (SUPPORT VECTOR MACHINES - SVM)

Según *Hernández, J. 2004*, afirma que las máquinas de vectores soporte pertenecen a los clasificadores lineales puesto que inducen separadores lineales o hiperplanos en espacios de características de muy alta dimensionalidad (introducidos por funciones núcleo o kernel) con un sesgo inductivo muy particular (maximización del margen).

En primer lugar, recordemos que todo hiperplano en un espacio D-dimensional, RD, se puede expresar como $h(x) = \langle w, x \rangle + b$, donde $w \in$ RD es el vector ortogonal al hiperplano, $b \in$ RDy $\langle \cdot, \cdot \rangle$ expresa el producto escalar habitual en RD.

Visto como un clasificador binario, la regla de clasificación se puede expresar como: $f(x) = signo(h(x))$, donde la función signo se define como:

$$signo(x) = \begin{cases} +1 \; si \; x \geq 0 \\ -1 \; si \; x < 0 \end{cases}$$

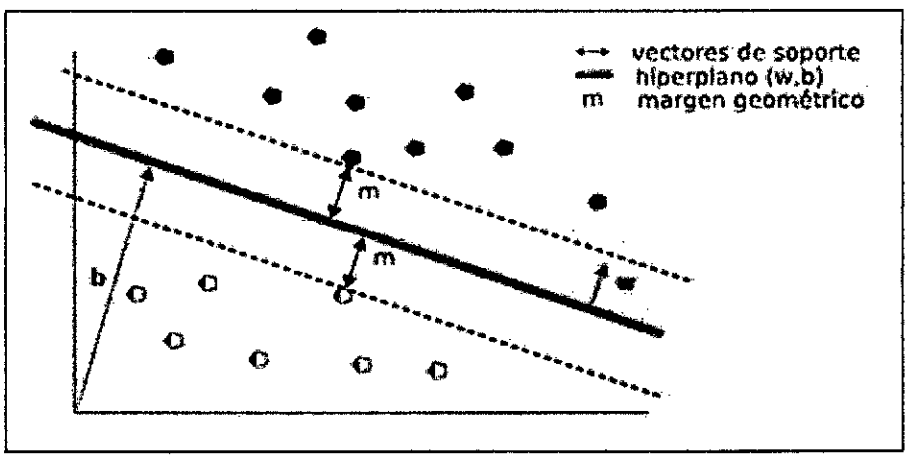
En la terminología de clasificación, las $x \in$ RD son representaciones vectoriales de los ejemplos, con una componente real por cada atributo y el vector *w* se suele denominar *vector de pesos*. Este vector contiene un peso para cada atributo indicando su importancia o contribución en la regla de clasificación. Finalmente, *b* suele denominarse sesgo (bias) y define el umbral de decisión. Dado un conjunto binario (es decir, con dos clases) de datos (ejemplos, vectores o puntos) linealmente separables, existen diversos algoritmos incrementales (on-line) para construir hiperplanos *(w, b)* que los clasifiquen correctamente.

Podemos citar, por ejemplo: *Perceptron, Widrow-Hoff, Winnow, Exponentiated - Gradient, Sleeping Experts,* etc. A pesar de que esté

garantizada la convergencia de todos ellos hacia un hiperplano solución, las particularidades de cada algoritmo de aprendizaje pueden conducirnos a soluciones ligeramente distintas, puesto que puede haber varios (de hecho infinitos) hiperplanos que separen correctamente el conjunto de ejemplos.

Suponiendo que el conjunto de ejemplos es linealmente separable, ¿Cuál es el "mejor" hiperplano separador en términos de generalización? La idea que hay detrás de la SVM de margen máximo consiste en seleccionar el hiperplano separador que está a la misma distancia de los ejemplos más cercanos de la cada clase. De manera equivalente, es el hiperplano que maximiza la distancia mínima (o *margen geométrico*) entre los ejemplos del conjunto de datos y el hiperplano. Intuitivamente, este hiperplano está situado en la posición más neutra posible con respecto a las clases representadas por el conjunto de datos, sin estar sesgado, por ejemplo, hacia la clase más numerosa. Además, sólo considera los puntos que están en las fronteras de la región de decisión, que es la zona donde puede haber dudas sobre a qué clase pertenece un ejemplo (son los denominados *vectores soporte*).

Gráfico N°4: Hiperplano separador obtenido con la MVS



En la Figura se presenta geométricamente este hiperplano equidistante (o de margen máximo) para el caso bidimensional.
Este sesgo inductivo de aprendizaje consiste en maximizar el margen se justifica dentro de la teoría del aprendizaje estadístico y se enmarca en el *principio* de *Minimización de Riesgo Estructural* [Vapnik 1995]. En esta teoría, maximizar el margen geométrico se ve como una buena heurística para minimizar la "complejidad" de la clase de hiperplanos separadores, que, a su vez, interviene directamente en las expresiones que acotan superiormente el error de generalización. A nivel práctico, el hiperplano separador de margen máximo ha demostrado una muy buena capacidad de generalización en numerosos problemas reales, así como una robustez notable frente al sobreajuste u *overfitting.*

A nivel algorítmico, el aprendizaje de las SVM representa un problema de optimización con restricciones que se puede resolver usando técnicas de programación cuadrática (QP). La convexidad garantiza una solución única (esto supone una ventaja con respecto al modelo clásico de redes neuronales) y las implementaciones actuales permiten una eficiencia razonable para problemas reales con miles de ejemplos y atributos.

El aprendizaje de separadores no lineales con SVM se consigue mediante una transformación no lineal del espacio de atributos de entrada (*input space)* en un espacio de características (*featurespace)* de dimensionalidad mucho mayor y donde sí es posible separar linealmente los ejemplos. El uso de las denominadas *funciones núcleo* (*kernelfunctions*), que calculan el producto escalar de dos vectores en el espacio de características, permite trabajar de manera eficiente en el espacio de características sin necesidad de calcular explícitamente las transformaciones de los ejemplos de aprendizaje. El aprendizaje en espacios de características vía transformaciones no lineales por medio de funciones núcleo no exclusiva del paradigma SVM. Aunque se suele asociar los métodos basados en

funciones núcleo con las SVM, al ser su ejemplo más paradigmático y más avanzado, hay muchos otros algoritmos que pueden "kernelizarse" para permitir el aprendizaje de funciones no lineales. Éste es el caso, por ejemplo, del perceptrón, de los discriminantes de Fisher, del análisis de componentes principales, etc., métodos que se han visto en otros capítulos de este libro.

Un requisito básico para aplicar con éxito las SVM a un problema real es la elección de una función núcleo adecuada, que debe reflejar el conocimiento a priori sobre el problema. El desarrollo de funciones núcleo para estructuras no vectoriales (por ejemplo, estructuras secuenciales, árboles, grafos, etc.) es actualmente una importante área de investigación con aplicación en dominios como el procesamiento del lenguaje natural y la bioinformática, entre otros.

A veces los ejemplos de aprendizaje no son linealmente separables ni tan siquiera en el espacio de características. Otras veces no es deseable conseguir un separador perfecto del conjunto de aprendizaje, puesto que los datos de aprendizaje no están libres de errores (ejemplos mal etiquetados, valores de atributos mal calculados, inconsistencias, etc.), comportamientos excepcionales (*outliers*), etc. Focalizarse demasiado en todos los ejemplos de aprendizaje puede comprometer seriamente la generalización del clasificador aprendido por culpa del sobreajuste. En estos casos es preferible ser más conservador y admitir algunos ejemplos de aprendizaje mal clasificados a cambio de tener separadores más generales y prometedores. Este comportamiento se consigue mediante la introducción del modelo SVM con margen blando (*softmargin*). En este caso, la función objetivo a minimizar está compuesta por la suma de dos términos: el margen geométrico y un término de regularización que tiene en cuenta los ejemplos mal clasificados. La importancia relativa de los términos se regula mediante un parámetro, normalmente llamado C. Este modelo, aparecido en 1995, es el que realmente abrió la puerta a un uso real y práctico de las SVM, aportando robustez frente al ruido.

Elementos matemáticos

Hay l observaciones y cada una consiste en un par de datos:

Un vector $x_i \in R^n$, i = 1,..., l

Una etiqueta $y_i \in \{+1, -1\}$

Supóngase que se tiene un hiperplano que separa las muestras positivas (+1) y las muestras negativas (-1). Los puntos x_i que están en el hiperplano satisfacen w*x+b=0.

Donde:

w es normal al hiperplano.

|b|/||w|| es la distancia perpendicular del hiperplano al origen.

||w|| es la norma euclidea de w.

Lo que se quiere es separar los puntos de acuerdo al valor de su etiqueta y_i en dos hiperplanos diferentes:

Hiperplano positivo: w*x+b =+1

Hiperplano negativo: w*x+b = -1

CASOS

- **Caso linealmente separable**

Cada punto de entrenamiento $X_i \in R^N$ pertenece a alguna de dos clases y se le ha dado una etiqueta $Y_i \in \{-1,1\}$ para i = 1, ..., l.

Una solución a esta situación es mapear el espacio de entrada en un espacio de características de una dimensión mayor y buscar el hyperplano óptimo.

Sea, z = $\varphi(x)$ la notación del correspondiente vector en el espacio de características con un mapeo φ de R^N a un espacio de característica Z.

Se muestra el hyperplano en la siguiente función:

$$w * z + b = 0$$

Definido por (w, b), tal que se separe el punto x_i de acuerdo a la función:

$$f(x_i) = \text{sign}(w * z_i + b) = \begin{cases} 1 & yi = 1 \\ -1 & yi = -1 \end{cases}$$

Donde:

$w \in Z$ y $b \in R$.

Por consiguiente, el conjunto S se dice que es linealmente separable si existe (w,b), tal que las inecuaciones:

$$\begin{cases} (w * z_i + b) \geq 1 & , \ y_i = 1 \\ (w * z_i + b) \leq -1 & , \ y_i = -1 \end{cases} \qquad i = 1, \ldots, l$$

Sean válidas para todos los elementos del conjunto S. Para el caso linealmente separable de S, podemos encontrar un único hyperplano óptimo, para el cual, el margen entre las proyecciones de los puntos de entrenamiento de dos diferentes clases es maximizado.

- **Caso no linealmente separable.**

Si el conjunto S no es linealmente separable, surgen violaciones a la clasificación deben ser permitidas en la formulación de la SVM.

Para tratar con datos que no son linealmente separables, el análisis previo puede ser generalizado introduciendo algunas variables no-negativas $\varepsilon_i \geq 0$ de tal modo que es modificado a:

$$y_i(w * z_i + b) \geq 1 - \varepsilon_i, \ i=1,\ldots, l$$

$$\text{s.a}\sum_{i=1}^{l} y_i \alpha_i = 0,\ 0 \leq \alpha_i \leq C\ ,\ i=1,\ldots,l$$

y la función de decisión es la siguiente:

$$f(x) = sign(w * z + b) = sign(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x_j) + b)$$

2.5.2.3. Redes neuronales

Se debe conocer cuál fue su origen y evolución en el transcurso del tiempo. Los primeros teóricos que concibieron los fundamentos de la computación neuronal fueron Beltran Russell, Warren McCulloch y Walter Pitts, quienes en 1943 lanzaron una teoría, acerca de la forma de trabajar de las neuronales. Ellos modelaron una red neuronal simple mediante circuitos eléctricos. Entre los años 1940 y 1950, los científicos comenzaron a pensar seriamente en la Red Neuronal utilizando como concepto la noción de que las neuronas del cerebro funcionan como interruptores digitales (on-off) de manera similar se desarrolló el computador digital, así nace la idea de la revolución cibernética.

En 1956, el Congreso de Dartmouth indico el nacimiento de la inteligencia artificial. Frank Rosenblatt comenzó el desarrollo del Perceptron en 1957, este modelo era capaz de generalizar, es decir, después de haber aprendido una serie de patrones podía reconocer otros similares, aunque no se le hubiesen presentado en el entrenamiento. Sin embargo, tenía una serie de limitaciones, por ejemplo, su incapacidad para resolver el problema de la función OR exclusiva y era incapaz de determinar clases no separables linealmente. En su libro confirmó que bajo ciertas condiciones el aprendizaje del Perceptron convergía hacia un estado finito. Bernard Widroff y MarcianHoff desarrollaron los modelos Adaline (ADAptative LINear Elements) y Madaline (Múltiple ADALINE) en 1960, estos fueron aplicados a un problema real de filtros adaptativos para eliminar ecos en las líneas telefónicas. Paul Werbos desarrolló la idea básica del algoritmo de aprendizaje de *propagación hacia atrás* (back propagation) en 1974; cuyo significado quedó definitivamente aclarado

en 1985. A partir de 1986, el panorama fue alentador con respecto a las investigaciones y el desarrollo de las redes neuronales. En 1988 fue formada la Sociedad Internacional de Redes Neuronales. En la actualidad, son numerosos los trabajos que se realizan y publican cada año, las aplicaciones nuevas que surgen (sobre todo en el área de control) y las empresas que lanzan al mercado productos nuevos, tanto hardware como software (sobre todo para simulación).

2.5.2.3.1. Definición de la Red Neuronal

Las Redes Neuronales o Redes Neuronales Artificiales (RNA) son técnicas de explotación de datos, es decir extrae datos de una información implícita, desconocida y potencialmente útil. Las Redes Neuronales comprenden muchos modelos y métodos de aprendizaje. Una red neuronal consiste en un modelo de nodos e interconexiones que recrea el diseño de las interconexiones neuronales del cerebro humano. Respecto al modo interno de trabajo las redes neuronales son modelos matemáticos multivariantes que utilizan procedimientos iterativos, con el objetivo de minimizar una determinada función de error. Con las redes neuronales se puede realizar automática y eficientemente múltiples tareas como: modelación, optimización, *regresión, clasificación*, lógica difusa, patrones y rasgos ocultos, memorización, aprendizaje asociativo, control adaptativo, *Pronóstico y Predicción de Series de Tiempo*, etc. Los diferentes tipos de RNA son sistemas definidos por funciones denotados por $f(\cdot)$. Un sistema matemáticamente definido es una transformación que en forma única traza un patrón de entrada en un patrón de salida. Como se muestra en la *figura1*, cuando la entrada al sistema es denotada por el vector X y la salida denotada por el vector Y, la relación entrada-salida puede ser escrita como $Y=f(X,W)$, donde W denota los pesos de la red. Los pesos y la estructura de los nodos interconectados en el sistema definen la transformación de entrada=salida desarrollado por la red.

2.5.2.3.2. Neuronas biológicas y artificiales

La neurona recibe información a través de la sinapsis de sus dendritas. Cada sinapsis representa la unión un axón de otra neurona con una dendrita de la neurona representada en la figura. Una transmisión electro-química tiene lugar en la sinapsis, la cual permite a la información ser transmitida desde una neurona a la próxima. La información es entonces transmitida a lo largo de las dendritas hasta que alcanza el cuerpo de la cédula. Allí tiene lugar e sumatorio de los impulsos eléctricos que lo alcanzan y se aplica algún tipo de función de activación a éste. La neurona se activará si el resultado es superior a un determinado límite o umbral. Esto significa que enviará una señal (en forma de onda de ionización) a lo largo de su axón con la finalidad de comunicarse con otras neuronas. Ésta es la manera en la que la información pasa de una parte de la red de neuronas a otra. Es muy importante tener en cuenta que las sinapsis tienen diferente rendimiento y que éste cambia al transcurrir el tiempo de vida de la neurona.

Gráfico N°5: Diagrama Representativo de una neurona real

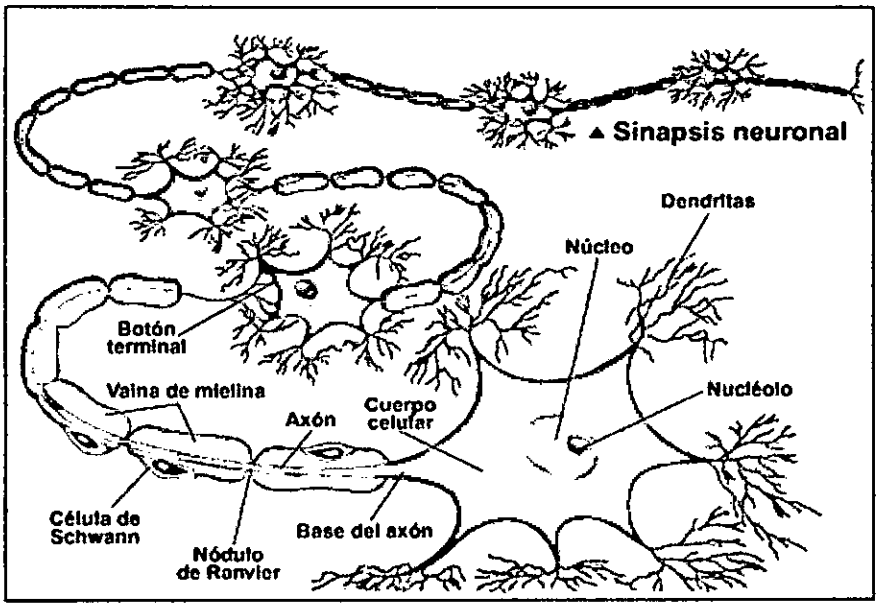
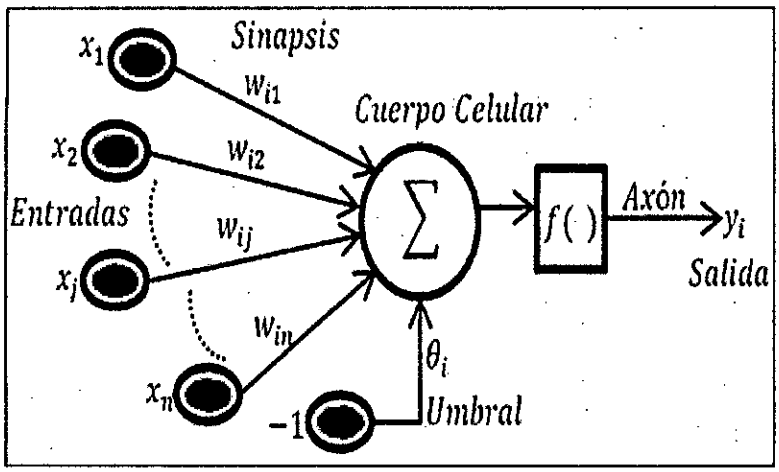


Gráfico N°6: Diagrama de una neurona artificial



Normalmente modelamos una neurona biológica de la manera que se muestra en la Figura (derecha). Esta figura incluye una entrada externa adicional, denominada polarización o "bias" y denotada por θ_i cuya finalidad es la de poder aumentar o disminuir el umbral de excitación de la neurona dependiendo de si es un valor positivo o negativo, respectivamente.

Las entradas se representan por el vector de entrada, x, y el rendimiento de las sinapsis se modela mediante un vector de pesos, ω. Entonces el valor de salida de esta neurona viene dado por:

$$y = f(\sum_i w_i x_i) = f(w, x) = f(w^T x)$$

Donde f es la función de activación.

Cuando tenemos una red de neuronas, las salidas de unas se conectan con las entradas de otras. Si el peso entre dos neuronas conectadas es positivo, el efecto es de inhibición.

Por tanto, podemos ver que una única neurona es una unidad de procesamiento muy simple, Se considera que el potencial de las redes neuronales artificiales

2.5.2.3.3. El aprendizaje en las redes neuronales artificiales

Hay dos tipos principales de aprendizaje en RNA:

- **Aprendizaje supervisado**. Con este tipo de aprendizaje, proporcionamos a la red un conjunto de datos de entrada y la respuesta correcta. El conjunto de datos de entrada y la respuesta correcta. El conjunto de datos de entrada es propagado hacia adelante hasta que la activación alcanza las neuronas de la capa de salida. Entonces podemos comparar la respuesta calculada por la red con aquella que se desea obtener, el valor real, objetivo o "blanco" (de *target,* en inglés). Entonces se ajustan los pesos para asegurar que la red produzca de una manera más probable una respuesta correcta en el caso de que se vuelva a presentar el mismo o similar patrón de entrada. Este tipo de aprendizaje será útil especialmente para las tareas de regresión y clasificación.

- **Aprendizaje no supervisado**. Sólo se proporciona a la red un conjunto de datos de entrada. La red debe auto-organizarse (es decir, auto enseñarse) dependiendo de algún tipo de estructura existente en el conjunto de datos de entrada. Típicamente esta estructura suele deberse a redundancia o agrupamientos en el conjunto de datos. Este tipo de aprendizaje será útil especialmente para las tareas de agrupamiento y reducción de dimensionalidad.

Al igual que otros paradigmas de la Inteligencia Artificial, la faceta más interesante del aprendizaje no es sólo la posibilidad de que los patrones de entrada puedan ser aprendidos/ clasificados/ identificados sino la capacidad de generalización que posee. Es decir, mientras el aprendizaje tiene lugar en un conjunto de patrones de entrenamiento, una propiedad importante de éste es que la red pueda generalizar sus resultados en un conjunto de patrones de prueba los cuales no han sido vistos durante el aprendizaje. Uno de los problemas a tener en cuenta es el peligro de sobreaprendizaje, denominado más técnicamente "sobreajuste".

- Aprendizaje supervisado en RNA

Para introducir este tipo de aprendizaje primero presentamos dos de las primeras redes neuronales que lo emplearon en su diseño y posteriormente mostraremos dos de las redes neuronales más usadas basadas en la utilización de éste.

2.5.2.3.4. Métodos para la evaluación de las redes neuronales

Para el aprendizaje con redes neuronales se debe considerar los siguientes parámetros:

- Seleccionar el tipo de función de entrada y salida.
- El número de nodos en la capa de entrada. Es definido por el número de variables independientes o predictoras consideradas en el modelo.
- El número de capas ocultas. Se puede determinar en forma automática con el algoritmo. Otras opciones son: (Nr.Atributos + Nr.Clases)/2, Nr.Atributos + Nr.Clases, Nr.Atributos o Nr.Clases.
- El número de nodos en la capa de salida. Es definido por el número de valores posibles que tiene el atributo clase.

2.5.2.3.5. Perceptrón simple y adaline

El perceptrón simple fue inicialmente investigado por Rosenblatt en 1962 [Rosenblatt, 1962]. El perceptrón simple tiene una estructura de varios nodos o neuronas de entrada y uno o más de salida. Un perceptrón simple, por tanto, no tiene capa oculta y así su estructura es como la red neuronal artificial anterior, pero sin ninguna capa oculta o intermedia.

Asociado a un patrón de entrada particular, x^P, tenemos una salida o^P y un "blanco" o salida correcta t^P. El algoritmo tiene la siguiente forma:

1. La red comienza en un estado aleatorio. Los pesos entre neuronas poseen valores pequeños y aleatorios (entre -1 y 1).

2. Seleccionar un vector de entrada, x^P, a partir del conjunto de ejemplos de entrenamiento.
3. Se propaga la activación hacia delante a través de los pesos en la red para calcular la salida $o^P = w.x^P$.
4. Si $o^P = t^P$ (es decir, si la salida de la red es correcta) volver al paso 2.
5. En caso contrario el cambio de los pesos se realiza atendiendo a la siguiente expresión: $\Delta w_i = \eta x_i^P (t^P - o^P)$ donde η es un número pequeño positivo conocido como coeficiente de aprendizaje. Volver al paso 2.

Lo que se hace, por lo tanto es ajustar los pesos de una manera en la que las salidas de la red, o^P , se vayan haciendo cada vez más semejantes al valor de los blancos, t^P, a medida que cada entrada, x_P, se va presentando a la red.

Otra red neuronal importante fue la de Adaline (*ADAptativeLINearElement*), concebida por Widrow y sus colaboradores en 1960 [Widrow&Hoff 1960]. Su topología es idéntica al perceptrón simple, es decir, no tiene capa oculta, pero la red Adaline calcula sus salidas empleando la siguiente expresión:

$$o = \sum_j w_j x_j + \theta$$

Con la misma notación de antes. La diferencia entre esta red y el Perceptrón es la presencia no de un umbral, θ. El interés en esta red se debió parcialmente al hecho de que se puede implementar fácilmente empleando un conjunto de resistores o interruptores.

La suma del error cuadrático a partir del uso de esta red en todos los patrones de entrenamiento viene dada por la siguiente expresión:

$$E = \sum_P E^P = \frac{1}{2} \sum_P (t^P - o^P)^2$$

Y el incremento de los pesos viene dado por su gradiente:

$$\Delta_P w_j = -\gamma \frac{\partial E^P}{\partial w_j}$$

Donde γ representa el coeficiente de aprendizaje. Esta regla se denomina Error Cuadrático Medio (*Least Mean Square Error,* LMS) o regla Delta o de Widrow-Hoff.

Ahora, en el caso del modelo Adaline con una sola salida, o, tenemos:

$$\frac{\partial E^P}{\partial w_j} = \frac{\partial E^P}{\partial o_P} \frac{\partial o^P}{\partial w^P}$$

Y debido a la linealidad de las unidades Adaline,

$\frac{\partial o^P}{\partial w_j} = x_j^P$ y también: $\frac{\partial E^P}{\partial o^P} = -(t^P - o^P)$.

Por tanto:

$$\Delta_P w_j = \gamma (t^P - o^P) x_j^P$$

Nótese la similitud entre esta regla de aprendizaje y la del perceptrón. Sin embargo, esta regla tiene mayor aplicación ya que se puede usar tanto para neuronas binarias como continuas, es decir, tanto para neuronas cuyas salidas son solamente ceros y unos o aquellas cuya salida son números reales. Es una de las reglas más potentes y se emplea como base de muchos métodos que utilizan aprendizaje supervisado.

El perceptrón simple y el modelo de Adaline son redes sin capa intermedia y, por tanto, si ignoramos las funciones de activación, son equivalentes a una función discriminante lineal.

2.5.2.3.6. Perceptrón multicapa

Tanto el Perceptrón y el modelo Adaline son métodos potentes de aprendizaje, aunque hay algunas situaciones en las que no dan lugar a buenos resultados. Estos casos se caracterizan por ser no linealmente separables. Hoy en día es posible mostrar que muchos conjuntos de datos que no son linealmente separables pueden ser modelados mediante el empleo del Perceptrón Multicapa (*MultilayerPerceptron,* MLP), es decir una red neuronal en forma de cascada, que tiene una o más capas ocultas, como se puede observar en la figura anterior.

Aunque esta potencialidad del MLP se descubrió pronto, se tardó bastante tiempo en encontrar un método o regla de aprendizaje apropiada para construirlas a partir de ejemplos. Esta regla parece que fue descubierta de manera independiente varias veces, y no existe acuerdo de la fecha exacta ni de su descubridor, pero fue popularizada principalmente por el GRUPO PDP (*Paralel Distributed Procesin*) [McClelland et al. 1986], bajo el nombre de Retropropagación o Propagación hacia atrás, que veremos más adelante.

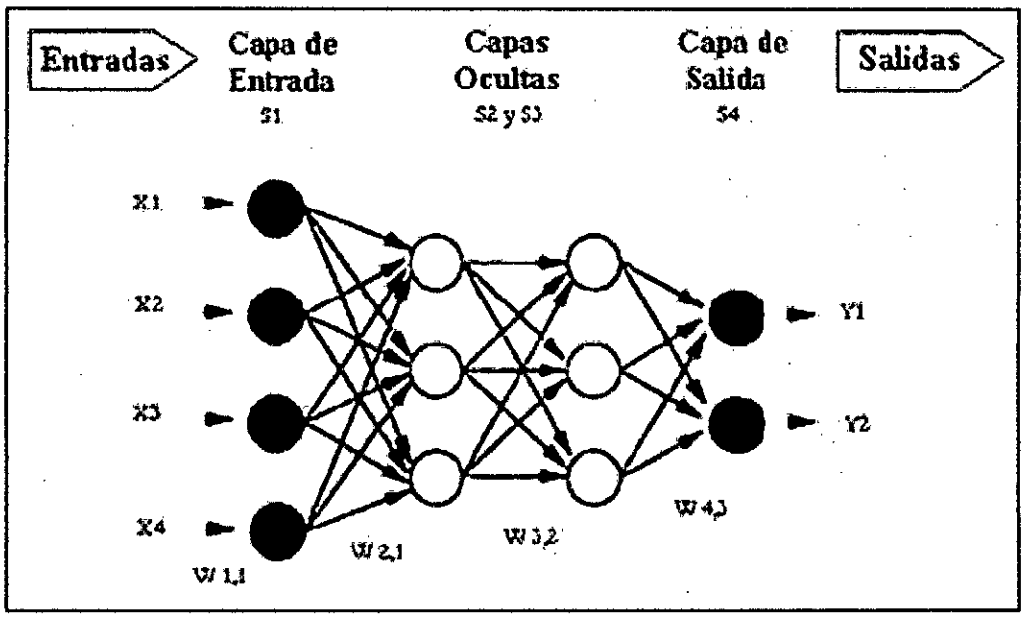
Respecto al uso de la red o de la activación, la activación se propaga en la red a través de los pesos desde la capa de entrada hacia la capa intermedia donde se aplica alguna función de activación a las entradas que le llegan. Entonces la activación se propaga a través de los pesos hacia la capa de salida.

Por tanto, si pensamos en el aprendizaje, hay que actualiza dos conjuntos de pesos: aquellos entre la capa oculta o intermedia y la de salida, y aquellos entre la capa de entrada y la capa intermedia. El error debido al primer conjunto de pesos se calcula empleando el método del error cuadrático medio anterior descrito. Entonces se propaga hacia atrás la parte del error debido a los errores que tienen lugar en el segundo conjunto de pesos y se asigna el error proporcional a los pesos que lo causan.

Podemos utilizar cualquier número de capas ocultas que queramos ya que el método es bastante general. Sin embargo, un factor a tener en cuenta es normalmente el tiempo de entrenamiento, el cual puede ser excesivo para

arquitecturas con muchas capas. Además, se ha demostrado que redes con una única capa oculta son capaces de aproximar cualquier función continua (o incluso cualquier función con sólo un número finito de discontinuidades), en el caso de utilizar funciones de activación diferenciables (no lineales) en la capa oculta.

Grafico N°8: Estructura de un Percentrón multicapa



2.5.2.3.7. Algoritmo de retropopagación

Se observan los siguientes pasos para el algoritmo respectivo:

1. Inicializar los pesos a valores pequeños aleatorios.
2. Escoger un patrón de entrada, x^P, y presentarlo a la capa de entrada.
3. Propagar la activación hacia delante a través de los pesos hasta que la activación alcance las neuronas de la capa de salida.
4. Calcular los valores de "δ" para las capas de salida $\delta_j^P = (t_j^P - o_j^P) f'(Act_j^P)$ usando los valores de los blancos deseados para el patrón de entrada seleccionado.

5. Calcular los valores de "δ" para la capa oculta usando $\delta_i^P = \sum_{j=1}^{N} \delta_j^P w_{ji} f'(Act_i^P)$.
6. Actualizar los pesos de acuerdo con: $\Delta_P w_{ij} = \gamma \delta_i^P o_j^P$.
7. Repetir del paso 2 al 6 para todos los patrones de entrada.

2.5.2.4. Regresión logística

La Regresión Logística (RL), es una técnica que permite estudiar la dependencia funcional entre una variable dependiente categórica dicotómica (dos clases) o politómica (más de dos clases) y un conjunto de variables independientes o predictoras que pueden ser cuantitativas o categóricas. La RL como TMD (técnica de minería de datos), se aplica a los problemas de clasificación o predicción.

Montgomery, P. 2010, Regresión Logística. Se considerará el caso en el que la variable de respuesta, en un problema de regresión, sólo asume dos valores posibles: 0 y 1; esos números podrían ser asignaciones arbitrarias a una respuesta cualitativa.

$$y_i = x_i'\beta + \varepsilon_i$$

En donde $x_i' = [1, x_{i1}, x_{i2}, \dots, x_{ik}]$, $\beta' = [\beta_0, \beta_1, \beta_{2,\dots,}\beta_k]$, y la variable de respuesta toma los valores 0 o 1. Se supondrá que la variable de respuesta y_i es una variable aleatoria de Bernoulli, cuya distribución de probabilidad es la siguiente:

y_i	Probabilidad
1	$P(y_i = 1) = \pi_i$
0	$P(y_i = 0) = 1 - \pi_i$

Ahora bien, como $E(\varepsilon_i) = 0$, el valor esperado de la variable respuesta es

$$E(y_i) = 1(\pi_i) + 0(1 - \pi_i)$$

$$= \pi_i$$

Esto implica que

$$E(y_i) = x_i'\beta = \pi_i$$

Que quiere decir que la respuesta esperada, determinada con la función de respuesta $E(y_i) = x_i'\beta$ no es más que la probabilidad de que la variable de respuesta tenga el valor de 1.

Hay algunos problemas sustantivos con el modelo de regresión. El primero es que se observa que si la respuesta es binaria, entonces los términos de error ε_i sólo pueden tener dos valores, que son

$$\varepsilon_i = 1 - x_i'\beta \qquad cuando\ y_i = 1$$

$$\varepsilon_i = -x_i'\beta \qquad cuando\ y_i = 0$$

En consecuencia, no es posible que los errores en este modelo sean normales. En segundo lugar, la varianza del error no es constante, ya que

$$\sigma_{y_i}^2 = E\{y_i - E(y_i)\}^2$$

$$= (1 - \pi_i)^2\pi_i + (0 - \pi_i)^2(1 - \pi_i)$$

$$= \pi_i(1 - \pi_i)$$

Obsérvese esta última expresión equivale a

$$\sigma_{y_i}^2 = E(y_i)[1 - E(y_i)]$$

Porque $E(y_i) = x_i'\beta = \pi_i$, lo que indica que la varianza de las observaciones (que es igual a la varianza de los errores, porque $\varepsilon_i = y_i - \pi_i$, y π_i es constante) es una

función de la media. Por último hay una restricción para la función de respuesta, ya que

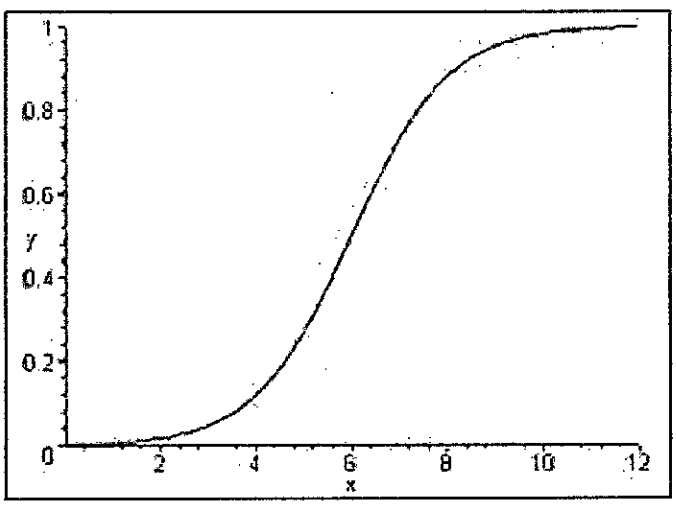
$$0 \leq E(y_i) = \pi_i \leq 1$$

Esta restricción puede causar graves problemas en la elección de una función de repuesta lineal, como se ha supuesto al principio. Sería posible ajustar al modelo con los datos para los cuales los valores predichos de la respuesta salen del intervalo 0, 1.

En general, cuando la variable de respuesta es binaria, hay bastantes pruebas empíricas que indican que la forma de la función de respuesta sebe ser no lineal. Una función monótonamente creciente (o decreciente), en forma de *S* (o de *S* invertida), es la que se acostumbra emplear: esta función se llama función de respuesta logística y tiene la forma

$$E(y) = \frac{\exp(x'\beta)}{1 + \exp(x'\beta)}$$

Gráfico N°9: Función de Regresión Logística



Función de respuesta logística. $E(y) = 1/(1 + e^{-6.0-1.0x})$

O bien, lo que es igual

$$E(y) = \frac{1}{1 + \exp(-x'\beta)}$$

La función de respuesta logística se puede linealizar con facilidad. Un enfoque cosiste en definir la porción estructural del modelo en términos de una función de la media de la función respuesta. Sea

$$\eta = x'\beta$$

El predictor lineal, estando definida η por la transformación

$$\eta = ln\frac{\pi}{1-\pi}$$

A esta transformación se le llama con frecuencia transformación logit de la probabilidad π, y la relación $\pi/(1-\pi)$ en la transformación se llama ventaja; a veces, a la transformación logit se le llama ventaja logarítmica.

Hay otras funciones que tienen la misma forma que la función logística, y también se puede obtener transformando π, una de ellas es la transformación probit, obtenida transformando a π con la distribución normal acumulada. De esta manera se obtiene un modelo de regresión probit, este modelo es menos flexible que el de regresión logística, y es probable que no se use tanto, porque no puede incorporar con facilidad más de una variable de π, definida por $\ln[-\ln(1-\pi)]$, que produce una función de respuesta que no es simétrica respecto al valor $\pi = 0.5$.

2.5.2.4.1. Estimación de parámetros en un modelo de regresión logística

La forma general del doble modelo de regresión logística es:

$$y_i = E(y_i) + \varepsilon_i$$

Donde las observaciones y_i son variables aleatorias independientes de Bernoulli, cuyos valores esperados son

$$E(y_i) = \pi_i$$

$$= \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}$$

Se usará el método de máxima verosimilitud para estimar los parámetros del predictor lineal $x_i'\beta$.

Cada observación de la muestra sigue la distribución de Bernoulli, por lo que la distribución de probabilidades de cada observación es

$$f_i(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \qquad i = 1,2, \dots, n$$

Y naturalmente, cada observación y_i toma el valor 0 o 1. Como las observaciones son independientes, la función de verosimilitud no es más que

$$L(y_1, y_2, \dots, y_n, \beta) = \prod_{i=1}^{n} f_i(y_i)$$

$$= \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

Es más cómodo trabajar con el algoritmo de la verosimilitud:

$$\ln L(y_1, y_2, \dots, y_n, \beta) = \ln \prod_{i=1}^{n} f_i(y_i)$$

$$= \sum_{i=1}^{n} [y_i ln\,(\frac{\pi_i}{1 - \pi_i})] + \sum_{i=1}^{n} \ln(1 - \pi_i)$$

Ahora bien, como $1 - \pi_i = [1 + \exp(x_i'\beta)]^{-1}$, y $\eta_i = \ln[\pi_i/(1 - \pi_i)] = x_i'\beta$, el logaritmo de la verosimilitud se puede expresar como sigue:

$$\ln L(y,\beta) = \sum_{i=1}^{n} \ln[1 + \exp(x_i'\beta)]$$

Con frecuencia, en los modelos de regresión logística se tienen observaciones o intentos repetidos en cada nivel de las variables x, esto sucede mucho en los experimentos diseñados. Sea y_i la cantidad de 1 observado en i, y n_i la cantidad de intentos en cada observación, entonces, el logaritmo de la verosimilitud se transforma en

$$\ln L(y,\beta) = \sum_{i=1}^{n} y_i\pi_i + \sum_{i=1}^{n} n_i \ln(1-\pi_i) - \sum_{i=1}^{n} y_i \ln(1-\pi_i)$$

Se podrían usar métodos numéricos de búsqueda, para calcular los estimados $\hat{\beta}$ por máxima verosimilitud (MLE, por *máximum likelihoodestimates*); sin embargo, sucede que se pueden usar los mínimos cuadrados iterativamente reponderados (IRLS) para determinar los MLE.

Sea $\hat{\beta}$ el estimado final de los parámetros del modelo que se obtiene con el algoritmo anterior. Si son correctas las hipótesis del modelo, se puede demostrar que, en forma asintótica.

$$E(\hat{\beta}) = \beta \qquad y \qquad Var(\hat{\beta}) = (X'^{V^{-1}}X)^{-1}$$

El valor estimado del predictor lineal es $\hat{\eta}_i = x_i'\hat{\beta}$, y el valor esperado del modelo de regresión logística se escribe con frecuencia como sigue:

$$\hat{y}_i = \hat{\pi}_i = \frac{\exp(\hat{\eta}_i)}{1 + \exp(\hat{\eta}_i)}$$

$$= \frac{\exp(x_i'\hat{\beta})}{1 + \exp(x_i'\beta)}$$

$$= \frac{1}{1 + \exp(-x_i'\beta)}$$

2.5.2.4.2. Pruebas de hipótesis para los parámetros del modelo

La prueba de hipótesis en la regresión logística (y en general, para el modelo lineal general) se basa en pruebas en pruebas de cociente de máxima verosimilitud, que es un procedimiento para muestras grandes, por lo que los procedimientos de prueba se basan en la teoría asintótica. El método de la razón de verosimilitud conduce a un estadístico llamado desviación.

Desviación del modelo

La desviación del modelo compara el logaritmo de la verosimilitud del modelo ajustado con el logaritmo de la verosimilitud de un modelo saturado, que es un modelo que tiene exactamente *n* parámetros y se ajusta perfectamente a los datos de la muestra. Para el modelo de regresión logística eso significa que las probabilidades de π_i son totalmente irrestrictas, por lo que al igualar $\pi_i = y_i$ (recuérdese que $y_i = 0$ *o* 1 se maximizaría la verosimilitud. Se puede demostrar que esto da como resultado el valor máximo de la función verosimilitud para el modelo saturado de unidad, por lo que el valor máximo de la función logaritmo de verosimilitud es cero.

Ahora se examinará la función logaritmo de verosimilitud para el modelo logístico ajustado. Cuando los estimados $\hat{\beta}$ de máxima verosimilitud se usan en la función logaritmo de verosimilitud, ésta alcanza su valor máximo, el cual es:

$$\ln L(\hat{\beta}) = \sum_{i=1}^{n} y_i x_i' \hat{\beta}_i - \sum_{i=1}^{n} \ln[1 + \exp(x_i' \hat{\beta})]$$

El valor de la función logaritmo de verosimilitud, para el modelo ajustado, nunca podrá ser mayor que el de esa función para el modelo saturado, porque el modelo ajustado contiene menos parámetros. La desviación compara al algoritmo de verosimilitud del modelo saturado con el algoritmo de verosimilitud del modelo ajustado. En forma específica, la desviación del modelo se define como sigue:

$$\lambda(\beta) = 2\ln L(modelo\ saturado) - 2\ln L(\hat{\beta})$$

$$= 2[\mathcal{L}(modelo\ saturado) - \mathcal{L}(\hat{\beta})]$$

Donde $\mathcal{L}$ representa el logaritmo de la función verosimilitud. Ahora bien, si el modelo de regresión logística es la función correcta de regresión, y el tamaño *n* de la muestra es grande, la desviación del modelo tiene aproximadamente una distribución ji cuadrada, con n-p grados de libertad. Valores grandes de la desviación del modelo indican que el modelo no es correcto, mientras que un valor pequeño implica que el modelo ajustado (que tiene menos parámetros que el modelo saturado) se ajusta a los datos casi tan bien como el modelo saturado. Los criterios formales de prueba son los siguientes:

Si $\lambda(\beta) \leq \chi^2_{\alpha,n-p}$ se concluye que el modelo ajustado es adecuado

Si $\lambda(\beta) > \chi^2_{\alpha,n-p}$ se concluye que el modelo ajustado no es adecuado

La desviación está relacionada con una cantidad muy conocida. Si se considera el error normal estándar del modelo de regresión lineal, sucede que la desviación es el error de la suma de cuadrados de residuales dividido entre la varianza del error σ^2.

Prueba de hipótesis sobre subconjuntos de parámetros usando la desviación

También se puede usar la desviación para probar hipótesis sobre subconjuntos de los parámetros del modelo, tal como se usaron diferencia de (o error de) las sumas de cuadrados, para probar hipótesis en el caso del modelo de regresión lineal con errores normales. Recuérdese que el modelo se puede escribir en la forma:

$$\eta = X\beta$$

$$= X_1\beta_1 + X_2\beta_2$$

Donde el modelo completo tiene *p* parámetros, β_1 contiene a *p-r* de esos parámetros, β_2 contiene a *r* de esos parámetros, y las columnas de las matrices X_1

y X_2 contienen las variables asociadas con esos parámetros. Supóngase que se desea probar la hipótesis:

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

Por consiguiente, el modelo reducido es:

$$\eta = X_1\beta_1$$

Ahora se ajusta el modelo reducido y se define a $\lambda(\beta_1)$ como la desviación para el modelo reducido, esta desviación siempre será mayor que la del modelo completo, porque el modelo reducido contiene parámetros; sin embargo, si la desviación del modelo reducido no es mucho más grande que la del modelo completo, quiere decir que el modelo reducido no es mucho más grande que la del modelo completo, quiere decir que el modelo reducido tiene un ajuste más o menos tan bueno como el modelo completo, por lo que es probable que los parámetros en β_2 probablemente no sea cero, y se debe rechazar la hipótesis nula. Formalmente, la diferencia en la desviación es:

$$\lambda(\beta_2|\beta_1) = \lambda(\beta_i) - \lambda(\beta)$$

Y esta cantidad tiene *n-(p-r)-(n-p)=r* grados de libertad. Si es cierta la hipótesis nula y si *n* es grande, la diferencia de la desviación tiene una distribución ji cuadrada con *r* grados de libertad. Por consiguiente, el estadístico de prueba y los criterios de decisión son

$$\text{Si } \lambda(\beta_2|\beta_1) \geq \chi^2_{\alpha,r} \text{ rechazar la hipótesis nula}$$

$$\text{Si } \lambda(\beta_2|\beta_1) < \chi^2_{\alpha,r} \text{ no rechazar la hipótesis nula}$$

A veces, la diferencia de desviación Si $\lambda(\beta_2|\beta_1)$se llama desviación parcial, que es una prueba de cociente de verosimilitud. Para visualizarlo, sea $L(\hat{\beta})$ el valor máximo de la función verosimilitud para el modelo completo, y sea $L(\widehat{\beta_1})$ el valor máximo de la función de verosimilitud para el modelo reducido. El cociente de verosimilitud es:

$$\frac{L(\widehat{\beta_1})}{L(\hat{\beta})}$$

El estadístico para la prueba de cociente de verosimilitud es igual a -2 multiplicado por el logaritmo del cociente de verosimilitud, es decir:

$$\chi^2 = -2ln\frac{L(\widehat{\beta_1})}{L(\hat{\beta})}$$

Sin embargo, es exactamente igual que la diferencia de desviación.

Pruebas de los coeficientes individuales del modelo

Se pueden hacer pruebas de los coeficientes individuales del modelo, como

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

Aplicando el método de la diferencia de la desviación. Hay otro método que también se basa en la teoría de los estimadores de máxima verosimilitud. Para muestras grandes, la distribución de un estimador de máxima verosimilitud es aproximadamente normal, con poco o ningún sesgo, además, las varianzas y covarianzas de un conjunto de estimadores de máxima verosimilitud se pueden determinar a partir de las segundas derivadas parciales de la función logaritmo de verosimilitud, con respecto a los parámetros del modelo, evaluadas en los estimados de máxima verosimilitud, entonces se puede hacer un estadístico *t* para probar las hipótesis de arriba. A esto a veces se le llama inferencia de Wald.

Sea G la matriz de *pxp* de las segundas derivadas parciales de la función logaritmo de verosimilitud, esto es,

$$G_{ij} = \frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_i \partial \beta_j}, \qquad i, j = 0, 1, \ldots, k$$

G se llama matriz hessiana o de Hess. Si los elementos de la hessiana se evalúan en los estimadores de máxima verosimilitud $\beta = \hat{\beta}$, la matriz de covarianza para muestra grande, de los coeficientes de regresión, es

$$Var(\hat{\beta}) = \hat{\Sigma} = -G(\hat{\beta})^{-1}$$

Las raíces cuadradas de los elementos diagonales de esta matriz son los errores estándar de muestras grandes de los coeficientes de regresión, por lo que el estadístico de prueba para la hipótesis nula en

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

Es:

$$Z_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

La distribución de referencia para este estadístico es la distribución normal estándar. Algunos programas de cómputo elevan al cuadrado el estadístico Z_0 y lo comparan con una distribución ji cuadrada con un grado de libertad.

CAPÍTULO III: METODOLOGÍA EMPLEADA

3.1. MÉTODOS EMPLEADOS EN LA INVESTIGACIÓN

En este apartado se han utilizado 4 métodos la cual se detalla en el marco conceptual:

Arboles de decisión

Redes Neuronales

Máquinas de Soporte Vectorial (SVM)

Regresión Logística

3.2. METODOLOGÍA PARA LA PRUEBA DE HIPÓTESIS

Para comprobar la hipótesis se utilizaron indicadores de comparación de modelos como la Validación Cruzada Generalizada (GCV), la tabla de confusión y la Curva COR.

1) Se obtuvieron los diferentes modelos de aprendizaje de máquina ajustados para la misma base de datos.
2) Se obtuvo el modelo clásico de creditscoring ajustado para los mismos datos
3) Mediante los indicadores de comparación mencionados anteriormente se hace un ranking de los modelos con mejor poder predictivo
4) Los que estén mejores en el ranking serán considerados los modelos con mejor poder predictivo.

3.3. TÉCNICAS E INSTRUMENTOS EMPLEADOS

En la presente investigación, se procedió a utilizar la base de datos otorgada por una entidad financiera para la aplicación de diferentes técnicas de creditscoring, la cual se mantiene en reserva el nombre de la institución, por temas de seguridad y prestigio de la entidad financiera.

3.4. PROCEDIMIENTO DE LA RECOLECCIÓN DE DATOS

El procedimiento de recolección de los datos fue de manera interna por la entidad financiera antes mencionada, la cual consta de las siguientes variables:

Cuadro N°3: Descripción de variables

Variables	Descripción	Valores
X1	Nivel de ingreso económico	Alto, Medio, Bajo
X2	Número de tarjetas	Menos de 5, 5 o más
X3	Nivel educativo	Bachilleratos, Universitario
X4	Número de carros	1 ó ninguno, 2 ó más
X5	Edad	Mínimo=20, máximo=64
Y	Valor de crédito	Bueno, Malo

El número de clientes que consta la base de datos es de 2464 clientes. De los cuales es 70% pertenecen al aprendizaje (1725 clientes) y el 30% a la prueba (739 clientes).

CAPÍTULO IV: DESARROLLO DEL ANÁLISIS E INTERPRETACIÓN

4.1. ANÁLISIS, INTERPRETACIÓN Y DISCUSIÓN DE RESULTADOS

En este apartado se presentan los resultados obtenidos, acompañados de su respectiva explicación y un análisis profundo. Se utilizaron tablas y gráficas para reportar los resultados y así facilita su comprensión.

A continuación, presentaremos algunos resultados de los modelos considerados en la presente investigación:

En primer lugar, debemos señalar que se trabajó con el 70% de los datos para el **entrenamiento** de los modelos, la cual fue seleccionada mediante un muestreo aleatorio simple para la selección de la misma, posteriormente con la data restante se procedió a la **validación** de los modelos y así poder observar el poder predictivo de los modelos estimados.

4.1.1. REGRESIÓN LOGÍSTICA.

Para el modelo paramétrico nos otorga la siguiente estimación:

Cuadro N°4: Resultados del Desarrollo del Análisis e Interpretación

```
    Educacion + Creditos_coche, family = binomial(link = "logit"))

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.889445  -0.628111   0.195548   0.615580   2.566352

Coefficients:
                    Estimate   Std. Error   z value             Pr(>|z|)
(Intercept)      -3.393821011  0.343523698  -9.87944 < 0.0000000000000002 ***
Edad              0.107985162  0.008888172  12.14931 < 0.0000000000000002 ***
Ingresos          1.773396822  0.108571524  16.33390 < 0.0000000000000002 ***
Tarjetas_credito -2.453057603  0.243665556 -10.06731 < 0.0000000000000002 ***
Educacion         0.164137522  0.134464775   1.22067              0.22221
Creditos_coche    0.066850833  0.215048040   0.31086              0.75590
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2350.4860  on 1724  degrees of freedom
Residual deviance: 1399.7407  on 1719  degrees of freedom
AIC: 1411.7407
```

Modelo con interacción

Para el modelo paramétrico nos otorga la siguiente estimación:

Cuadro N°5: Modelo con interacción

```
Call:
glm(formula = Valoracion_credito ~ Edad + Ingresos + Tarjetas_credito +   Educacion + Credi
tos_coche + Ingresos * Tarjetas_credito +
    Ingresos * Educacion + Ingresos * Creditos_coche + Tarjetas_credito *  Educacion + Tarj
etas_credito * Creditos_coche + Educacion *
    Tarjetas_credito, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9234  -0.6134   0.2240   0.5909   2.5493

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -3.281786   0.419969  -7.814 5.52e-15 ***
Edad                           0.108975   0.008959  12.163  < 2e-16 ***
Ingresos                       1.478823   0.253890   5.825 5.72e-09 ***
Tarjetas_credito              -2.533154   0.506607  -5.000 5.73e-07 ***
Educacion                      0.571255   0.357018   1.600    0.110
Creditos_coche                -0.235308   0.570222  -0.413    0.680
Ingresos:Tarjetas_credito      0.255336   0.362821   0.704    0.482
Ingresos:Educacion            -0.175857   0.216996  -0.810    0.418
Ingresos:Creditos_coche        0.248356   0.322973   0.769    0.442
Tarjetas_credito:Educacion    -0.273968   0.338626  -0.809    0.418
Tarjetas_credito:Creditos_coche 0.036685   0.555772   0.066    0.947
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2350.5  on 1724  degrees of freedom
Residual deviance: 1394.9  on 1714  degrees of freedom
AIC: 1416.9

Number of Fisher Scoring iterations: 5
```

Claramente, las variables Educación y Crédito coche no son significativas a un nivel de confianza del 95% así como todas las interacciones propuestas, y el criterio de AIC nos otorga 1416, por ende proponemos el modelo inicial.
Claramente las variables Educación y Crédito coche no son significativas a un nivel de confianza del 95%, y el criterio de AIC nos otorga 1411.

En la estimación de los datos de **entrenamiento** nos otorgó la siguiente Matriz de confusión.

Cuadro N°6: Estimación logística

	Referencia	LOGISTICA	
Predicción		0	1
	0	552	149
	1	178	846

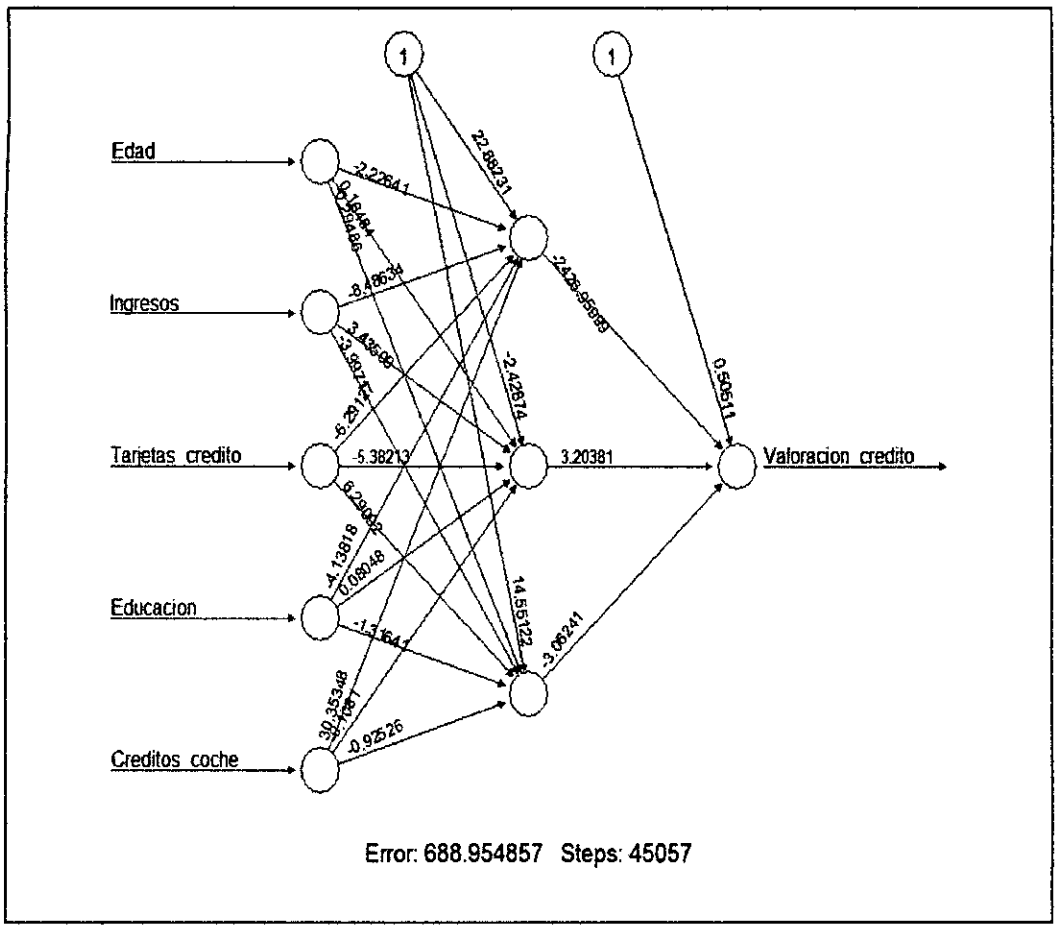
Elaboración: Propia

Como se puede observar en la matriz existen un total de 1398 casos correctamente clasificados para la muestra de entrenamiento la cual representa un 81.04% de porcentaje de acierto.

4.1.2. REDES NEURONALES ARTIFICIALES

Para el primer modelo no paramétrico, se utilizó las Redes Neuronales Artificiales, en la que se utilizó el método de retropropagación resiliente para la estimación de los pesos, la cual se muestra en el siguiente gráfico.

Gráfico N°10 Red Neuronal de la base de datos financiera



Elaboración: Propia

La red neuronal se utilizó 1 capa oculta con 3 neuronas, este bajo la premisa de obtener el mejor modelo parsimonioso, y con mayor poder predictivo, un aprendizaje de 0.01 por defecto.

En el gráfico de pesos generalizados para explorar la significancia de las variables, exploratoriamente la edad no era significativa para el modelo puesto en el gráfico se observa que la edad se encuentra en el cero, pero al estimar nuevamente el modelo se verifico que el error aumento por lo que se consideró todas las variables para una comparación de todos los modelos presentados en el presente estudio.

En la estimación de los datos de entrenamiento se obtuvo la siguiente tabla de matriz de confusión.

Cuadro N°7: Estimación de Redes Neuronales

Predicción	Referencia	**RNA**	
		0	1
	0	534	130
	1	196	865

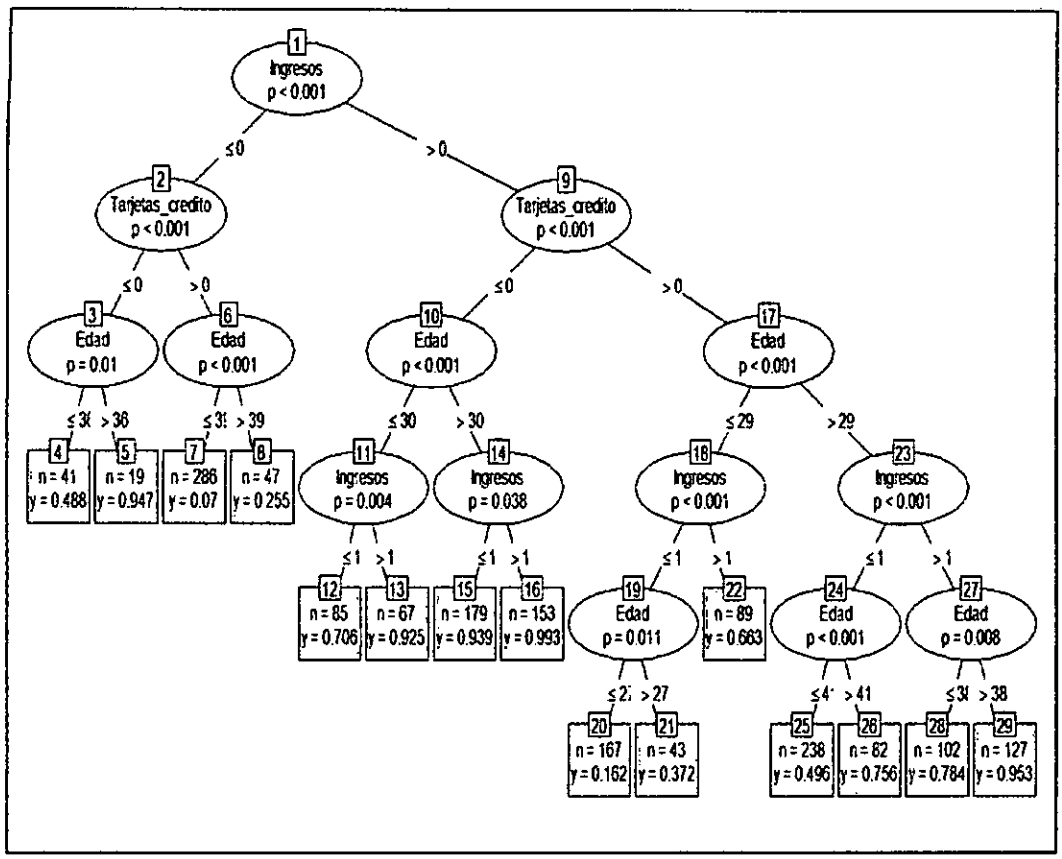
Elaboración: Propia

Como se puede observar en la matriz existen un total de 1399 casos correctamente clasificados para la muestra de entrenamiento la cual representa un 81.10% de porcentaje de acierto.

4.1.3. ÁRBOLES DE DECISIÓN

Para el tercer modelo no paramétrico se utilizó la técnica de clasificación de árboles de decisión, la cual nos otorga el siguiente grafico donde se observa las reglas de asociación y decisión.

Gráfico N°11: Reglas de decisión utilizando Arboles para la base de datos financiera



Elaboración: Propia

Como se puede observar en el grafico presenta las reglas de decisión a considerar en el modelo donde claramente utiliza las variables: ingresos, edad y tarjetas de créditos, las demás no son significativas para el modelo planteado.

En la estimación de los datos de entrenamiento se obtuvo la siguiente tabla de matriz de confusión.

Cuadro N°8: Estimación de Árboles de Decisión

Predicción	Referencia ARBOL		
		0	1
	0	609	213
	1	121	782

Elaboración: Propia

Como se puede observar en la matriz existen un total de 1391 casos correctamente clasificados para la muestra de entrenamiento la cual representa un 80.64% de porcentaje de acierto.

4.1.4. MÁQUINA DE SOPORTE VECTORIAL

Para el cuarto modelo no paramétrico planteado en el presente estudio, la cual se trabajó con todas las variables de estudios y utilizando un kernel lineal para la estimación puesto líneas abajo se podrá observar que tiene mejor porcentaje del área bajo la curva en las comparaciones de las curvas ROC.

En la estimación de los datos de entrenamiento se obtuvo la siguiente tabla de matriz de confusión.

Cuadro N°9: Estimación del SMV

Predicción	Referencia SVM		
		0	1
	0	579	151
	1	177	817

Elaboración: Propia

Como se puede observar en la matriz existen un total de 1396 casos correctamente clasificados para la muestra de entrenamiento la cual representa un 80.97% de porcentaje de acierto.

4.2. DISCUSIÓN DE LOS RESULTADOS

4.2.1. Comparación de los modelos propuestos

En este capítulo hablaremos sobre los resultados mostrados en el capítulo anterior, donde se pudo verificar que el modelamiento de una data financiera, por medio de 4 modelos propuestos como se muestra a continuación:

- Modelo Paramétrico, Regresión logística.
- Modelo No Paramétrico 1, Redes Neuronales Artificiales
- Modelo No Paramétrico 2, Arboles de decisiones.
- Modelo No Paramétrico 3, Máquina de Soporte Vectorial.

En el modelo paramétrico se utilizó la regresión logística en la cual se estimó el modelo con todas las variables para que sea comparativo con los demás modelos.

En los modelos no paramétricos se puede verificar que el que obtuvo mayor porcentaje de acierto es el de máquina de soporte vectorial, que obtiene un 81.51% de acierto.

Ahora no solo debemos comparar porcentajes de aciertos para los datos de entrenamientos sino también de los datos de validación que como se comentó se trata del 30% de la información de la base de datos.

4.2.2. Validación.

En la etapa de la validación se tomó en forma aleatoria el 30% de la información y así poder saber el poder predictivo de cada modelo propuesto en este trabajo de investigación, donde se obtuvieron los siguientes resultados:

1) Logística

En la validación de los datos se obtuvo la siguiente tabla de matriz de confusión.

Cuadro N°10: Validación de Logística

	Referencia	**LOGISTICA**	
Predicción		0	1
	0	226	64
	1	73	376

Elaboración: Propia

Como se puede observar en la matriz existen un total de 602 casos correctamente clasificados para la muestra de validación la cual representa un 81.46% de porcentaje de acierto.

2) Redes Neuronales Artificiales

En la validación de los datos se obtuvo la siguiente tabla de matriz de confusión.

Cuadro N°11: Validación de Redes Neuronales Artificiales

	Referencia	**RNA**	
Predicción		0	1
	0	219	71
	1	62	387

Elaboración: Propia

Como se puede observar en la matriz existen un total de 606 casos correctamente clasificados para la muestra de validación la cual representa un 82% de porcentaje de acierto.

3) Árbol de decisión

En la validación de los datos se obtuvo la siguiente tabla de matriz de confusión.

Cuadro N°12: Validación de Árbol de Decisión

	Referencia	ARBOL	
Predicción		0	1
	0	238	52
	1	97	352

Elaboración: Propia

Como se puede observar en la matriz existen un total de 590 casos correctamente clasificados para la muestra de validación la cual representa un 79.84% de porcentaje de acierto.

4) Máquina de soporte Vectorial

En la validación de los datos se obtuvo la siguiente tabla de matriz de confusión.

Cuadro N°13: Validación de Máquina de Soporte Vectorial

	Referencia	SVM	
Predicción		0	1
	0	230	60
	1	81	367

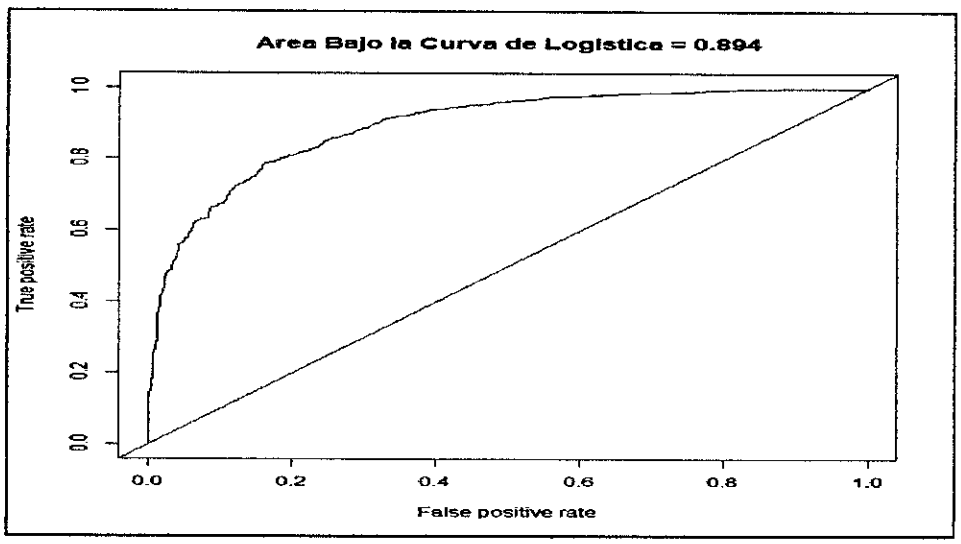
Elaboración: Propia

Como se puede observar en la matriz existen un total de 597 casos correctamente clasificados para la muestra de validación la cual representa un 80.89% de porcentaje de acierto.

4.2.3. CURVA ROC

Logística

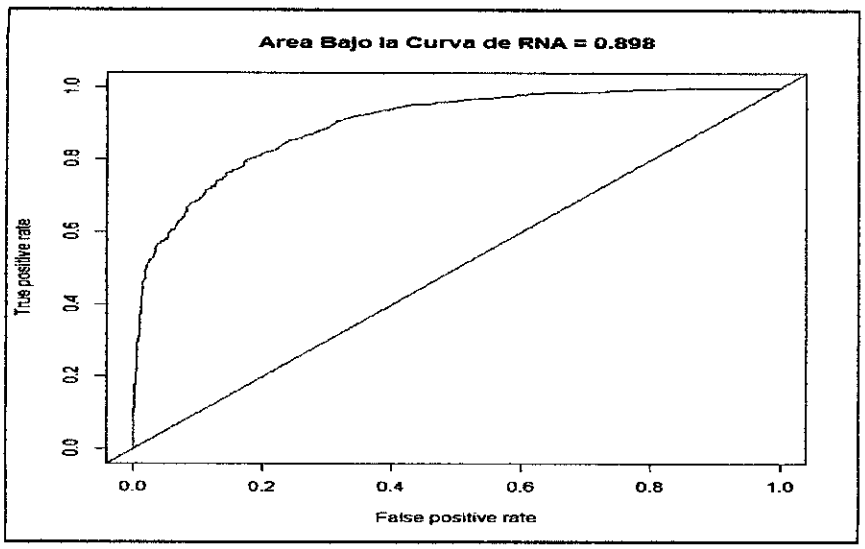
Gráfico N°12: Curva ROC – Logística



Elaboración: Propia

Redes Neuronales Artificiales

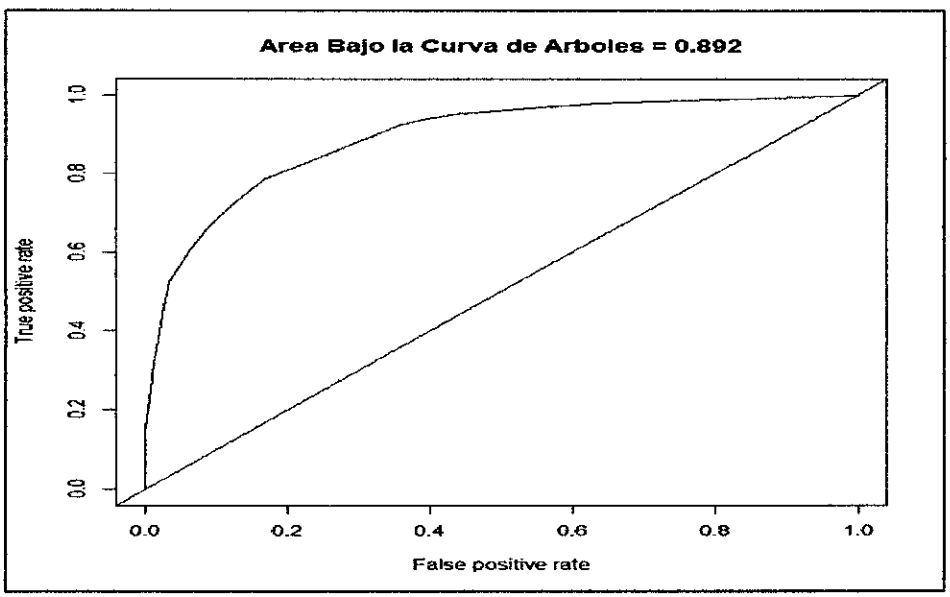
Gráfico N°13: Curva ROC – Redes Neuronales Artificiales



Elaboración: Propia

Arboles de decisión

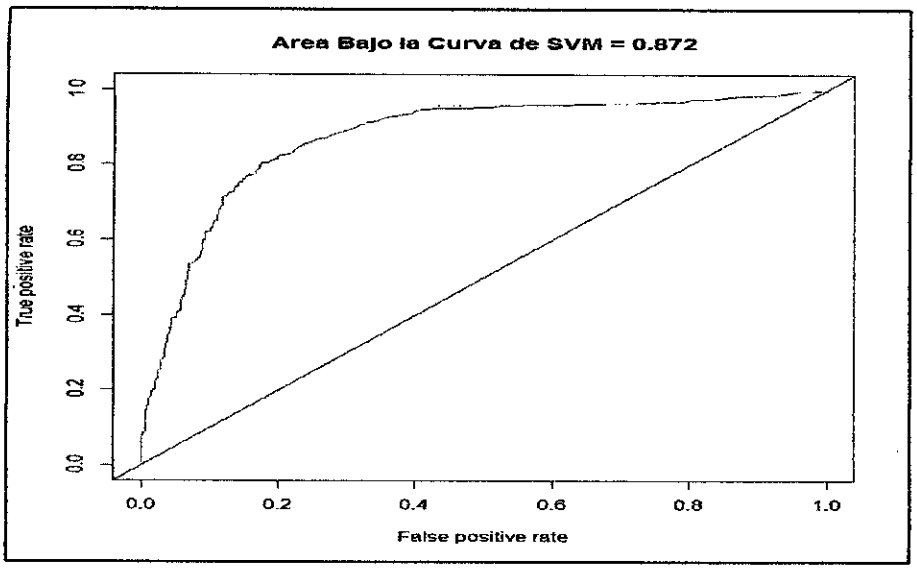
Gráfico N°14: Curva ROC – Árboles de decisión



Elaboración: Propia

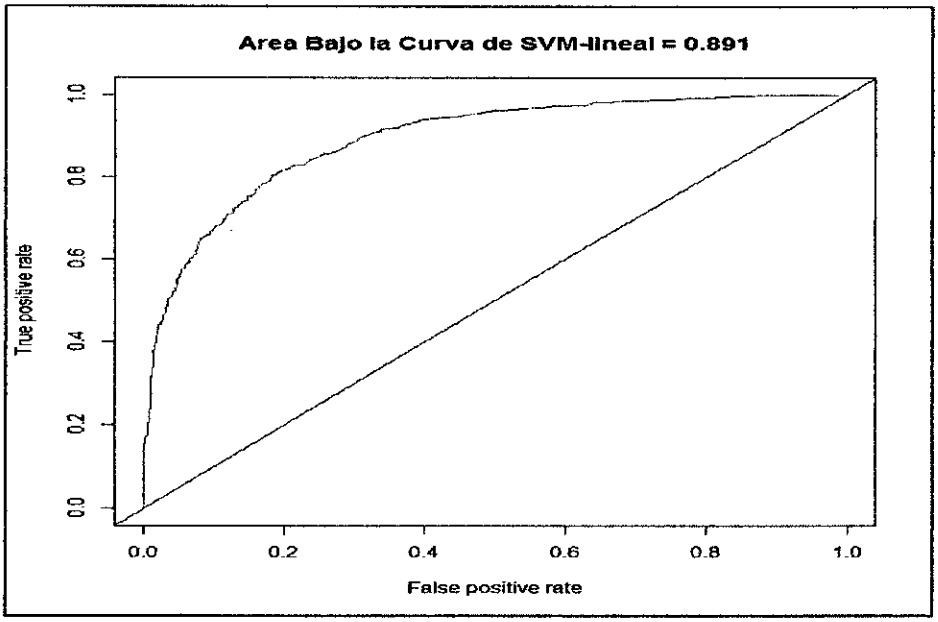
Máquina de soporte Vectorial

Gráfico N°15: Curva ROC – SMV Radial



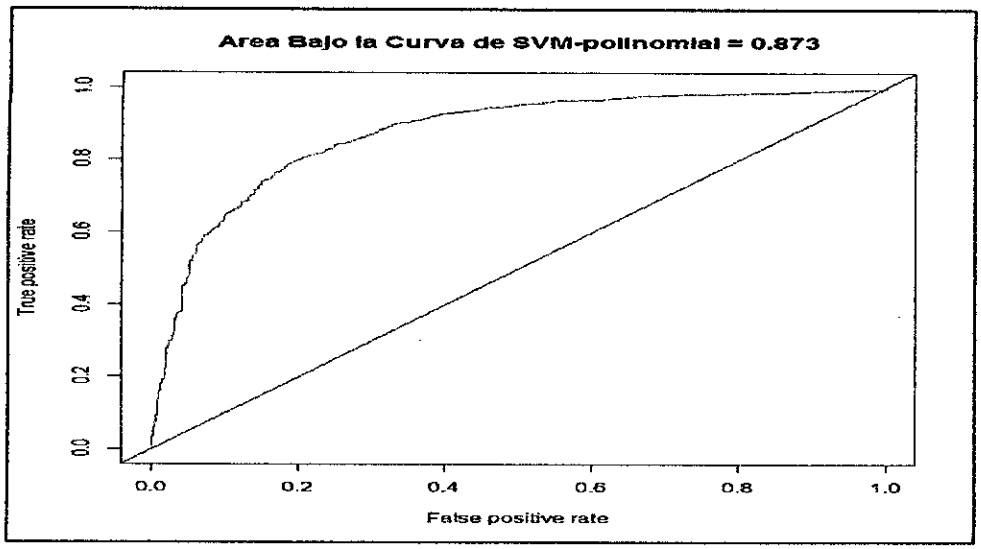
Elaboración: Propia

Gráfico N°16: Curva ROC – SMV Lineal



Elaboración: Propia

Gráfico N°17: Curva ROC – SMV Polinomial



Elaboración: Propia

Nota: Como se sabe el gráfico de curva ROC nos permite ver que tan bueno es nuestro modelo planteado con respecto a su predicción.

CAPÍTULO V: CONCLUSIONES Y SUGERENCIAS

5.1. CONCLUSIONES.

El riesgo crediticio estimado de personas naturales de una Institución Financiera de Chiclayo obtenido mediante los modelos de aprendizaje de máquina tales como Redes Neuronales, Máquinas de Soporte Vectorial, Árboles de Clasificación y el modelo clásico de Credit Scoring estimado mediante la Regresión Logística fueron 30.58%, 31.17%, 32.21%, 29.63% respectivamente.

Los modelos de aprendizaje de maquina estiman mejor el riesgo crediticio de personas naturales de una Institución Financiera de Chiclayo que el modelo de enfoque paramétrico, de la Regresión Logística para nuestros datos financieros. Por lo tanto los resultados de los modelos de aprendizaje de maquina presentados en la conclusión anterior presentan una mejor predicción de aciertos en la detección de clientes morosos o que no pagaran, objetivo de nuestra investigación considerado como el riesgo de que un cliente caiga en mora.

5.2. SUGERENCIAS

Asumiendo las conclusiones referidas en el subcapítulo anterior se muestra que el mejor modelo para esta investigación es la de redes neuronales con un 89.9% de poder predictivo, el cual se detalla en cuadro N° 12 donde se presentaron los AOC o el porcentaje de acierto bajo la curva. Por otro lado, el porcentaje de validación la cual se pronosticó fuera de muestra del método de redes neuronales frente a los demás modelos, se sobrepone a ellos con un ligero porcentaje de 82%.

Como primera sugerencia, dado que el acierto del riesgoso (0,0) es decir, del moroso es un 32.21%, una cifra considerable, para la toma de decisión del analista financiero en el momento de otorgar el crédito a los clientes, se podría considerar un mejor análisis documentario personal para que de esta manera la institución financiera cuente con una mejor garantía o respaldo al emitir su crédito.

Para terminar, esta investigación solo determina el modelo de pronóstico a elegir en torno a la toma de decisiones para el otorgamiento de crédito, sin embargo, en la actualidad ya se ha visto la necesidad de contar con un mejor patrón de conducta del cliente en especial los clientes de las cooperativas, uno de los factores puede ser por la cantidad inferior que sacan a crédito respecto a un banco, o tal vez sea por el déficit que tienen las personas que laboran en el trabajo de campo, es decir, a la hora de procesar la información brindada por el cliente, este hace filtrar u oculta algunos datos que son importantes para una adecuada toma de decisión por parte del analista financiero.

Con el sustento del párrafo anterior, una de las sugerencias tentativas es también darle una mejor instrucción o capacitación a los analistas financieros que acuden al campo, para que tengan mejor eficiencia y esto se plasme en beneficios para la entidad financiera.

BIBLIOGRAFÍA

Aguilar, G. (2004). *Análisis de la Morosidad de las Instituciones Microfinancieras(IMF) en el Perú.*Perú: Instituto de Estudios Peruanos.

Aldrich, J. H. (1984). *Linear probability.Logit and Probit models,sage publications.*USA.

Alfredo, R. M. (2002). *Aplicaciones estadísticas en la evaluación financiera de proyectos.* Colombia: Universidad Autónoma de Manizales.

Alvarado, M. (2002). *Evaluación y Manejo del riesgo Crediticio en el Ámbito Agrícola.* Perú.

Baca, G. A. (1997). *La Administración de Riesgos Financieros.* México: Revista Ejecutivos de Finanzas.

Bensic, M. S.-S. (2005). *Modelling small-business credit scoring by using logistic, neural networks and decision trees.*

Bodie, Z. y. (1999). *Finanzas.* México: Pretince Hall.

Cabrera, A. (2014). *Diseño de credit scoring para evaluar el riesgo crediticio en una entidad de ahorro y crédito popular.* México: Universidad Teconológica de la Mixteca.

Cardona Hernandez, P. A. (2004) Aplicación de árboles de decisión en modelos de riesgo crediticio. Colombia.

Escalona, A. (2011). *Uso de los modelos Credit Scoring en Microfinanzas.* México: Universidad Autónoma Chapingo.

Fragoso, J. (2002). *Análisis y Administración de Riesgos Financieros.* México.

Galicia Romero, M. (2003). *Nuevos Enfoques de riesgo de crédito.* México.

Hernandez, O. R. (2015). *Introducción a la minería de datos.*

Herrán, L. (2009). *Evaluación crediticia aplicando un modelo de credit scoring en el ámbito microempresaria.* Piura: Universidad de Piura.

Jorion, P. (1999). *Valor en riesgo.* México: Limusa.

Katare, A., & Athavale, V., (2011), Behavior Analysis of Different Decision Tree Algorithms. International Journal of Computer Technology and electronics Engineerring (IJICTEE) Vol. 1, Issue 1, pp. 43 - 47.

Krugman, R. P. (1995). *Economía Internacional, 3° Edición.* España: McGraw-Hill.

Ladino, I. (2014). *Comparación de modelos de riesgo de crédito: modelos logísticos y redes neuronales.* Bogotá, Colombia: Universidad Javeriana.

Levi, D. M. (1997). *Finanzas InternacionaleS, 3° Edición.*México: McGraw-Hill.

Lewent, J. C. (1990). *Identifying , Measuring and Hedging Currency Risk at Merck.* EEUU: Continental Bank Journal of Applied Corporate Finance 2.

Liao, T. F. (s.f.). *Interpreting Probability models,logit, probit and other generalizes linear models.* Sage Publications.

Long, J. (s.f.). *Regression models for categorical and limited dependet variables.*Sage Publications.

Mallo, F. (Julio 2011). *Modelos multivariantes internos de medición de riesgo de crédito, acordes con Basilea II.* Salamanca, España: Universidad de Salamanca.

Marzo, C.; Wicijowski, C.; Rodriguez, L. (2008). Prevención y cura de la morosidad. Análisis y evolución futura de la morosidad. España.

Moreno, S. (2013). *El Modelo Logit Mixto para la construcción de un Scoring de Crédito.* Colombia: Universodad Nacional de Colombia.

Nieto, S. (2010). *Crédito al consumo: La estadística aplicada a un problema de risego crediticio.* México: Universodad Autónoma Metropolitana.

Patil, N., Lathi, R., & Chitre, V. (2012). Customer Card Classification Based on C5.0 & CART Algorithms. International Journal of Engineering Research and Applications (IJERA), Vol.2, Issue 4, pp 164 - 167.

Peña, D. (2002). Análisis de datos multivariantes. España. Editorial: Interamericana de España

Pérez, C. (2011). Técnicas de segmentación. Conceptos, herramientas y aplicaciones. 1° Edición: Alfaomega Grupo Editor, S.A. de C.V., Mexico

Pérez, C. (2014). Técnicas estadísticas con variables categóricas. IBM SPSS. 1°Edición: Ibergarceta Publicaciones, S.L., Madrid.

Pérez, M. (2014). Mineria de datos a través de ejemplos. 1°Edicion: Alfaomega Grupo Editor, S.A. de C.V., Mexico.

Quilan, J. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.

Quilan, J. R. (1986). Induction of decision trees. Machine Learning, Vol. 1., pp. 86 - 106.

Resendiz Trejo, J. A. (2006). Las máquinas de vectores de soporte para identificar en linea. Mexico

Rosillo, J. (2002). *Modelo de predicción de quiebras de las empresas colombianas.* Colombia: Revista de Ciencias Administrativas y sociales, Universidad Nacional de Colombia.

Rodríguez, J. (2010). Fundamentos de minería de datos. Bogota 1° Edicion: Universidad Distrital Francisco José de Caldas.

Salinas Ávila, J. J. (2004). *Metodologías de Medición del Riesgo de Mercado en Instituciones de Fomento y Desarrollo Territorial.*Colombia.

Soltan, A. &. (1683-1688). *A hybrid model using decision tree and neural network for credit scoring problem.*Management Science Letters.

Valderrey, P. (2011). Segmentación de mercados. 1°edición: Ediciones de la U. Bogotá.

Véliz, C. (2016). Análisis multivariente: Métodos estadísticos multivariantes para la investigación. 1° Edición: Ciudad Autónoma de Buenos Aires: Cengage Learning Argentina.

Venkata Krishna Kumar, S. & Kiruthika, P. (2015). An Overview of Classification Algorithm in Data mining. International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 12, pp. 255.257

Vigo, G. (2010). *Método de clasificación para evaluar el riesgo crediticio: Una comparación.* Lima, Perú: Universida Nacional Mayor de San Marcos.

Anexos

Comparación de los indicadores de los modelos propuestos

Cuadro N°14: Tabla comparativa 1

Tabla Comparativa		
Técnica	**Estimación %**	**Validación %**
Logística	81.04	81.46
Redes neuronales	81.10	82.00
Árbol de decisión	80.64	79.84
Máquina de Soporte Vectorial	80.97	80.65

Fuente: Elaboración Propia

En la tabla comparativa de los modelos propuestos se observa en primer lugar la estimación otorgada para los datos de entrenamientos, la cual el 81.10% es el mayor valor, otorgada por las Redes Neuronales, asimismo este coincide con el poder predictivo fuera de muestra nos da un 82% de acierto, cabe recordar que el porcentaje de acierto es la suma de los pares (0,0) y (1,1), es decir el acierto de los riesgosos y no riesgosos reales y pronosticados por el modelo.

Ahora se muestra los resultados de los modelos que mejor predicción tiene acerca de los aciertos en la detección de clientes morosos o que no pagaran es decir el par (0,0), objetivo de nuestra investigación considerado como el riesgo de que un cliente caiga en mora.

Cuadro N°15: Tabla comparativa 2

Tabla Comparativa de los que aciertan que no pagaran		
Técnica	**Estimación %**	**Validación %**
Logística	30.96	29.63
Redes neuronales	32.00	30.58
Árbol de decisión	35.30	32.21
Máquina de Soporte Vectorial	33.58	31.17