(13)

Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA- IZTAPALAPA
DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERIA

MODELOS DE CALIFICACIÓN CREDITICIA: TÉCNICAS DE RECONOCIMIENTO DE PATRONES Y MODELOS ESTADÍSTICOS TRADICIONALES

Tesis que presenta
Adán Díaz Hernández

Para obtener el grado de
Maestro en Ciencias y Tecnologías de la Información

Asesor:	Dr. John Goddard Close	
Jurado Calificador:		
Presidente:	Dr. René Mackinney Romero	UAM-I
Secretario:	Dra. Patricia Saavedra Barrera	UAM-I
Vocal:	Dr. José Carlos Ramírez Sánchez	Anáhuac México-Norte

México, D.F. diciembre 2012

Resumen

En este trabajo se presenta una propuesta de modelos de calificación crediticia mediante el uso de técnicas de reconocimiento de patrones y modelos estadísticos tradicionales. La idea principal de este proyecto de investigación consiste en realizar una comparación del desempeño de los modelos estadísticos utilizados tradicionalmente en la industria financiera, con metodologías basadas en clasificadores desarrollados en el área de aprendizaje maquinal y reconocimiento de patrones. Esto se realiza bajo ejercicios de experimentación con distintos conjuntos de datos, tanto públicos como privados, en el contexto nacional como internacional.

Uno de los objetivos del trabajo se enfoca en realizar un análisis para el caso mexicano con información del mercado local que establezca las bases para el uso de alternativas cuyo desempeño sea competitivo, o incluso superior, respecto a los enfoques tradicionales empleados para discriminar el comportamiento y calidad crediticia de los clientes a quienes las instituciones otorgan crédito. Este problema se aborda desde una perspectiva estática (clasificación de buenos y

malos clientes en un punto en el tiempo), al tiempo que se sientan bases para futuros análisis dinámicos del comportamiento del nivel de riesgo del portafolio.

Agradecimientos

Con profundo agradecimiento a mis padres y hermanos, por el cariño y apoyo incondicional que siempre me han brindado, sin ellos nada de esto hubiera sido posible...

Un amplio reconocimiento para el Dr. John Goddard Close, director de esta tesis, cuyo invaluable apoyo, interés y dirección hicieron posible la culminación de esta investigación. Asimismo, quisiera hacer patente mi agradecimiento a la Dra. Patricia Saavedra, el Dr. René Mackinney y el Dr. José Carlos Ramírez por las valiosas aportaciones que hicieron para mejorar la presente investigación.

Con un enorme y especial cariño a Elisa. El camino recorrido para culminar este proyecto fue largo y pesado, y sin embargo, nunca dejaste de brindarme tu apoyo y comprensión.

A mi gran amigo Jesús, quien no dudó nunca en brindarme su invaluable ayuda cuando más la necesitaba.

También quisiera expresar mi agradecimiento a quienes estuvieron vinculados de alguna manera en este proyecto, en verdad he sido muy afortunado al conocer gente tan valiosa durante mi vida académica y profesional. Vienen a mi mente sus nombres y numerosos recuerdos, de manera que su apoyo y amistad incondicionales le dieron un sentido especial y único a esta última etapa como estudiante...

Gracias.

Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

El problema de calificación crediticia (CS, por sus siglas en inglés: *credit-scoring)* representa una de las primeras aplicaciones de la minería de datos (*data mining*), inclusive antes de que dicho término apareciera. Las instituciones de crédito tradicionalmente emplean metodologías estadísticas para discriminar a los clientes con buen perfil de riesgo de aquellos que potencialmente pueden presentar un mal perfil. Sin embargo, en los últimos años, diversas alternativas se han desarrollado en el área de aprendizaje maquinal (*machine learning*) para abordar este problema. El aprendizaje maquinal es una rama de la inteligencia artificial en la cual se diseñan y desarrollan algoritmos que permiten generar conclusiones sobre patrones a partir de datos empíricos. Los mecanismos de aprendizaje utilizan ejemplos (datos) para capturar características de interés de la distribución de probabilidad subyacente. Una de las líneas de investigación en aprendizaje maquinal se centra en el estudio del aprendizaje automatizado para reconocer patrones complejos y

realizar toma de decisiones sobre la base de cierto conjunto de datos. A esta área del aprendizaje maquinal se le denomina reconocimiento de patrones.

La capacidad de las instituciones financieras para pronosticar o anticipar la calidad de los portafolios de créditos que mantienen resulta un tema crucial para la continuidad de su negocio y más aún, para garantizar la estabilidad del sistema financiero y el crecimiento económico en su conjunto. Precisamente, la reciente crisis financiera global tuvo sus orígenes en los severos problemas financieros que enfrentó el sector hipotecario estadounidense, producto, entre otros factores, de una mala gestión del riesgo de crédito. Las deficiencias en los sistemas evaluación y calificación crediticia, tanto para contrapartes públicas como los acreditados en los distintos segmentos hipotecarios, condujeron no solamente a malas estimaciones de los riesgos financieros asociados sino a valuaciones erróneas de distintos activos y productos financieros derivados en el mercado[1]. El uso de adecuados sistemas de calificación crediticia es entonces un tema central para instituciones y reguladores a nivel local e internacional.

Típicamente, las aproximaciones empleadas en la industria financiera para la construcción de modelos de CS descansan fuertemente en metodologías estadísticas tradicionales como el análisis discriminante, la regresión logística, árboles de decisión, y, en algunos casos, la aplicación de redes neuronales artificiales. Al menos

[1] Algunos de los productos financieros que sufrieron impactos importantes por la mala valoración del riesgo de crédito fueron los respaldados por activos riesgosos (*asset-back securities*) y los derivados de crédito (*credit derivatives*). Las pérdidas que enfrentaron los activos respaldados por hipotecas residenciales llevaron a una crisis económica severa. Como referencia, el lector puede consultar el capítulo "Securitization and the Credit Crisis of 2007" en Hull (2012).

en el contexto del sector financiero local, el uso de técnicas alternativas como las desarrolladas en el campo de acción del reconocimiento de patrones es menos frecuente que el caso de técnicas estadísticas tradicionales como la regresión logística. Los esfuerzos de este proyecto se enfocan en explorar la aplicabilidad de tales metodologías sobre datos reales del mercado mexicano, evaluando las bondades y deficiencias encontradas, con la finalidad de sentar bases útiles para futuros desarrollos en la materia. Para fines comparativos con otros trabajos previos, en esta investigación se utilizan conjuntos de datos públicos, sobre los cuales se reportan los resultados obtenidos mediante el uso de los distintos clasificadores seleccionados.

En concreto, son cinco los conjuntos de datos que se emplean para los análisis empíricos. Tres de ellos (Alemán, Australiano y PAKDD) se encuentran disponibles de forma pública mientras que los dos restantes (Cerveza y Autos) son de tipo privado, cuyo origen no se proporciona por razones de confidencialidad. Los conjuntos Alemán y Australiano se encuentran disponibles en el repositorio de UCI (http://archive.ics.uci.edu/ml/) y han sido utilizados ampliamente en la literatura como *benchmark* para comparar distintas metodologías de clasificación y modelos de CS. Por su parte, el conjunto PAKDD se puede obtener mediante solicitud en la página de la 13th Pacific-Asia Knowledge Discovery and Data Mining conference (http://sede.neurotech.com.br:443/PAKDD2009/). El uso de información pública, la cual también ha sido empleada en otros estudios, obedece a la necesidad de

contar con un punto de comparación para los resultados obtenidos en este proyecto.

El conjunto Cerveza contiene información de clientes a quienes una empresa mexicana vendió cerveza (durante 2009 y 2010). Para una proporción de estos clientes, la venta del producto se realizó mediante el otorgamiento de líneas de crédito y se monitoreó el comportamiento de perfil de riesgo. Finalmente, el conjunto Autos contiene información socio-demográfica y de seguimiento de los patrones de pago de una cartera de préstamos. Los préstamos de este conjunto de datos fueron otorgados por una empresa mexicana especializada en el financiamiento automotriz durante el periodo comprendido del año 2006 al 2009. Los resultados obtenidos en un primer estudio estático de CS sientan las bases para realizar un estudio dinámico en el tiempo como futura línea de investigación.

Bajo el contexto de esta investigación, se entenderá por clasificador a toda técnica cuya tarea consista en construir una función $f: X \rightarrow C$, donde X denota al conjunto de valores que pueden tomar las características (o atributos) que describen a cierta muestra de observaciones (o ejemplos) y C es un conjunto formado por las distintas clases (o etiquetas), las cuales corresponden a los valores que puede tomar cierta variable de interés en el problema de estudio. Usualmente, la función f constituye un mapeo que asigna un valor específico $y \in C$ a cada configuración n-dimensional $x \equiv (x_1, x_2, \ldots, x_n) \in X$ que contiene valores específicos de los atributos. En el caso concreto del CS, el conjunto C que generalmente interesa cuando se utiliza como herramienta de otorgamiento (originación) de crédito está formado dos categorías:

aceptado/rechazado, en relación al resultado de aplicaciones de crédito de los clientes. Los atributos asociados a las aplicaciones corresponden a información socio-demográfica de los clientes (por ejemplo, edad, sexo, estado civil, ingresos mensuales, tipo de vivienda, etc) y demás información relevante para la evaluación del perfil de los clientes[2]. Asimismo, cuando el interés se centra en el riesgo asociado al incumplimiento de los clientes en el pago de sus obligaciones, se pueden definir las categorías del perfil de riesgo como bueno/malo, o bien, asignar etiquetas que indiquen las distintas clases de riesgo (pudiendo ser dos o más categorías). La definición de las categorías puede construirse sobre la base de algún sistema de calificaciones de riesgo (interno o externo), así como también algún otro criterio de interés como la rentabilidad y desempeño de los clientes a quienes se les ha otorgado algún crédito.

Como se ha mencionado antes, típicamente, entre las metodologías más utilizadas por las instituciones de crédito para calificar sus carteras de clientes, se encuentran el análisis discriminante lineal y la regresión logística. La fácil implementación y facilidad en la interpretación de los resultados hacen de estas dos técnicas las más socorridas en la industria. Los árboles de clasificación y las redes neuronales son algunos métodos alternativos a los que generalmente se recurre en la práctica (Thomas et al., 2002). En el caso concreto de México, los trabajos sobre CS que se

[2] En el caso en que la solicitud de crédito sea realizada por una persona moral (empresa) para solicitar algún tipo de financiamiento, los atributos comúnmente están relacionados con información de los estados financieros e indicadores que reflejen la capacidad de pago y salud financiera de la empresa.

pueden encontrar en la literatura son escasos. Altman (2005) aplicó una versión mejorada de su modelo Z-score para calificar compañías mexicanas, el cual es esencialmente un modelo lineal que incorpora razones financieras para la construcción de un puntaje (*score*). El que el modelo utilizado en empresas del mercado estadounidense resultó adaptable para el caso mexicano.

Objetivos específicos

En este documento se analizan diversas técnicas de clasificación para la construcción de modelos de CS. Entre las técnicas tradicionalmente empleadas, se encuentran el análisis discriminante, regresión lineal, árboles de clasificación y vecinos más cercanos. Asimismo, se estudian diversos métodos del área de reconocimiento de patrones para su aplicación en el contexto de los sistemas de calificación del crédito, entre las que destacan las redes neuronales, máquinas de soporte vectorial, clasificadores bayesianos (*naive bayes*, y redes bayesianas), así como modelos híbridos (*bagging* y *boosting*) y máquinas de Boltzmann restringidas[3]. Hasta donde se tiene conocimiento, esta última técnica no ha sido utilizada anteriormente en el contexto de CS, por lo que se en esta investigación se sentaría un precedente al respecto. Específicamente, se busca comparar la capacidad predictiva de los distintos clasificadores calibrados sobre los conjuntos de datos anteriormente descritos.

[3] Varias de las técnicas mencionadas se pueden revisar, por ejemplo, en Witten et al. (2011).

Los modelos implementados buscan analizar el perfil de riesgo, tanto de contrapartes que solicitan crédito, como de aquellos clientes previamente aceptados. Las características específicas de cada conjunto de información crediticia disponible permiten orientar el análisis bajo un contexto estático, al tiempo de plantear alternativas para implementar en el futuro estudios del comportamiento dinámico del perfil de riesgo de los clientes a lo largo del periodo de observación.

El documento está integrado por cinco capítulos. En el segundo capítulo se realiza una revisión de algunos aspectos básicos sobre el CS y su evolución histórica. Asimismo, se incluyen desde estudios iniciales sobre aproximaciones utilizadas para resolver el problema, hasta los últimos estudios encontrados sobre técnicas alternativas de clasificación empleadas. El tercer capítulo presenta distintas técnicas que han sido aplicadas para resolver problemas de calificación crediticia, las cuales provienen tanto del área de la estadística como de la inteligencia artificial y el reconocimiento de patrones. Este apartado provee al lector de una base sobre los fundamentos teóricos y características generales de distintos clasificadores, tomando en cuenta sus bondades y deficiencias como herramientas para la construcción de sistemas de calificación crediticia. En el capítulo cuatro se realiza una revisión de las distintas metodologías y métricas más utilizadas, tanto en la industria como en la academia, para evaluar el desempeño de los modelos de clasificación. La discusión se centra principalmente en los enfoques empleados en el área de CS. El capítulo cinco incluye un análisis descriptivo de las características de los conjuntos de datos utilizados, así como los resultados de la estimación de

cada uno de los modelos calibrados y la comparación de su desempeño en cada caso. Al final del documento se incluye un apéndice que contiene información descriptiva adicional de los conjuntos de datos disponibles.

Capítulo 2

Aspectos generales del *credit-scoring* y revisión de la literatura

En este capítulo se revisan ciertos aspectos relativos al *credit-scoring* (CS), entre los que se enlistan: el concepto general del crédito y CS, su historia, elementos regulatorios en la industria financiera (Acuerdos de Basilea). Asimismo, la última sección del capítulo contiene una revisión de la literatura relativa a distintas técnicas de clasificación que se han empleado en el contexto del CS. Para efectos de profundizar en algunos fundamentos teóricos y metodológicos subyacentes a las técnicas que se utilizan en esta investigación, se recomienda al lector remitirse al capítulo 3.

Entendemos el término *crédito*[4] como un préstamo sobre el que una de las contrapartes se compromete a devolver en tiempo y forma la cantidad solicitada

[4] Desde un punto de vista legal, el crédito es el derecho que una de las partes denominada acreedor tiene para obligar a otra (llamada deudor) a pagar en los términos y condiciones pactadas.

(más intereses, seguros y otros costos asociados, si los hubiera), el cual puede ser otorgado por una institución financiera a clientes que lo solicitan.

La información proporcionada en la solicitud se analiza mediante un proceso de valoración para decidir si se aprueba o no el otorgamiento del crédito al solicitante; la decisión depende de la información registrada del cliente (Bicer et al., 2010; Crook et al., 2007). Este proceso de valoración, conocido como calificación crediticia, es una disciplina que se ha desarrollado, y ha sido ampliamente adoptada, desde principios de 1960s. En su contexto más general, CS se define como el conjunto de modelos de decisión, y técnicas subyacentes, que ayudan a los prestamistas en su proceso de otorgamiento del crédito a los prestatarios. Estas técnicas no sólo ayudan a decidir a quién se le otorga el crédito, sino también el monto del mismo y las estrategias de operación que permitan fortalecer la rentabilidad del crédito (Thomas et al., 2002).

Los sistemas de calificación crediticia constituyen uno de los primeros usos que se le dio a la información relativa al comportamiento de los consumidores, mucho antes de que el término minería de datos emergiera como resultado de la fusión entre estadística, inteligencia artificial y aprendizaje maquinal. Su utilización por tiendas y comercios (*retailers*) y compañías de venta por catálogo se remonta a 1950s en EUA. Las metodologías desarrolladas en esa época, son ejemplo de las primeras herramientas utilizadas para administrar el riesgo de crédito.

En un entorno globalizado, caracterizado por el crecimiento acelerado en la demanda de servicios financieros cada vez más complejos y personalizados, las

lecciones aprendidas en las recientes crisis financieras y económicas han puesto de manifiesto a las instituciones financieras la importancia de contar con metodologías y modelos de medición de riesgos adecuados (Tapiero, 2010). En el caso específico del riesgo de crédito, y en particular, en lo relativo a los modelos de calificación crediticia, diversas técnicas de clasificación se han implementado para resolver el problema de discriminar adecuadamente los diferentes perfiles de riesgo de las contrapartes. Entre los métodos aplicados a la calificación crediticia de clientes, se encuentran el análisis discriminante de Fisher, regresión lineal, regresión logística, y más recientemente, algoritmos genéticos, redes neuronales artificiales, redes Bayesianas, máquinas de soporte vectorial (SVM), y métodos híbridos (Anderson, 2007; Bilgic et al., 2010; Chen et al., 2003; Hand et al., 1997a; Hand et al., 1997b; Huang et al., 2004; Huang et al., 2007; Shin et al. 2005; Tsai y Chen, 2010).

El uso de los sistemas y tecnologías de CS se ha extendido más allá de su propósito original de evaluar el riesgo de crédito de las nuevas solicitudes y portafolios de créditos. Actualmente, se utilizan también para evaluar la rentabilidad ajustada por riesgo de los créditos otorgados, establecer límites de crédito iniciales y futuros a los prestatarios, así como el apoyo a un rango amplio de los servicios relacionados como la detección y prevención de fraudes, sin olvidar por supuesto, el diseño de estrategias para mitigar las pérdidas y controlar los niveles de morosidad, cobranza y recuperaciones. Esto ha permitido promover la eficiencia y expandir el alcance de la cobertura de otorgamiento de crédito en el sistema financiero, permitiendo la

bancarización de sectores de la población que en el pasado no habían sido atendidos[5].

Bajo el contexto del sistema financiero y la economía global en su conjunto, el proceso de calificación crediticia se encuentra inmerso en una serie de mejores prácticas y buen gobierno corporativo[6]. Adicionalmente, los conceptos de ética de negocio y responsabilidad social representan un cambio significativo respecto al rol esperado de los negocios en la sociedad. Estos dos aspectos no influyen en la decisión de utilizar o no un modelo de CS particular, sino en la manera en cómo deben ser utilizados en beneficio de la sociedad[7].

Entre los aspectos legales relacionados con el otorgamiento de créditos, es importante considerar tanto el tipo de información que se puede utilizar en el proceso de calificación como su uso responsable y protección a la misma. Por ejemplo, los bancos se encuentran sujetos a altos estándares en materia de regulación para garantizar la protección y privacidad de los datos de sus clientes. Dependiendo de cada país, los principios mínimos básicos que se deben observar se

---

[5] Ver sección de citas al final del Apéndice.

[6] La primera se refiere a los procesos, técnicas, metodologías, y el uso de tecnología, equipo y recursos que han probado tener éxito como medios para alcanzar objetivos específicos de las instituciones. Por su parte, la buena gobernabilidad se refiere tanto al proceso de toma de decisiones como a los mecanismos por los que las decisiones son implementadas (o no implementadas), donde la rendición de cuentas de la administración a los accionistas y otras partes interesadas (*stakeholders*) debe asegurar una adecuada alineación de intereses.

[7] La ética de negocio se relaciona con la conducta de las instituciones en lo relativo a lo que es moralmente correcto, en tanto que la responsabilidad social comprende la conducta que tienen éstas hacia las necesidades, aspiraciones, y preocupaciones de la sociedad en su conjunto (se puede limitar simplemente a una conducta ética o incluso extender a la filantropía)

relacionan con la recolección de datos, razonabilidad de la información solicitada, su calidad, limitación en su uso, medidas de seguridad para su manejo y transmisión, revelación de la información (derecho al secreto bancario), acceso de clientes a su propia información y rendición de cuentas[8].

Otro principio relevante en la construcción de los modelos de CS se refiere a la prohibición del uso de campos de información discriminatoria. En la mayoría de los países se han implementado legislaciones anti-discriminatorias relacionadas con la "prohibición de la discriminación injusta", "promoción de igualdad de oportunidades" o "protección de los derechos humanos", las cuales aplican sobre contextos más generales que el crédito, como el empleo y otras prácticas. Las características que generalmente se prohíbe utilizar en los modelos de calificación son la raza, religión, nacionalidad y orientación sexual. En general, el consenso es que las características demográficas, sobre las que los consumidores no tienen control, debieran ser remplazadas por aquello que sí puedan cambiar, en particular comportamientos y estatus específicos a cada persona. Sin embargo, existen situaciones en las que algunas de estas características se pueden utilizar: (i) se carece de otro tipo de información crediticia; (ii) existe alguna justificación desde el punto de vista del negocio la cual puede ser probada; (iii) la información socio-demográfica utilizada es solamente un componente de una evaluación más amplia.

[8] En función de la regulación local, se pueden prever excepciones en cada caso, ya sea por consentimiento de los clientes, por requisición de la autoridad legal o bien por ser de interés público o nacional.

2.1 Acuerdos de Basilea

El Banco Internacional de Pagos (BIS, por sus siglas en inglés) es el banco central para la liquidación de transacciones internacionales entre bancos centrales. Este organismo promueve la cooperación en materia de supervisión bancaria y administración de riesgos financieros entre sus países miembros. En 1974 nace el Comité de Supervisión Bancaria de Basilea (BCBS, por sus siglas en inglés), integrado por los gobernadores de los bancos centrales del grupo de los diez (G-10). Para el año de 1988, producto de las lecciones aprendidas por crisis económicas y financieras en el pasado, 11 países suscriben el primer Acuerdo de Basilea (Basilea I) con al finalidad de establecer un marco de adecuación de capital. Basilea I fungió como un marco de referencia para la medición de la adecuación del capital de los bancos, así como para establecer los estándares mínimos que las autoridades supervisoras debían cumplir a fin de robustecer los mercados sujetos a su jurisprudencia.

El BCBS concluyó que una aproximación adecuada para determinar el capital es mediante una ponderación por riesgo de las diferentes categorías de activos; la ponderación se aplica dependiendo del riesgo relativo a una clasificación de los activos determinada por su calificación de riesgo. Dicha calificación de riesgo debía ser prestablecida por el regulador. A pesar de los beneficios que trajo consigo Basilea I, el acuerdo no contempló la cobertura del riesgo operacional ni la

flexibilidad para incorporar las innovaciones en la medición del riesgo[9]. Para tratar de solventar estos problemas, el BIS amplió en 1996 los acuerdos para incluir el riesgo de mercado, además publicó tres documentos consultivos (1999, 2001 y 2003), un estudio de impacto (2002) y varias publicaciones (2001) con el fin de adecuar el tratado a las condiciones cambiantes de mercado. Como resultado de esta nueva visión surgen los nuevos Acuerdos de Basilea (Basilea II). Además de la inclusión de un nuevo cargo de capital por riesgo operacional[10], Basilea II surge como un estándar para mejorar la medición de los riesgos y la asignación de capital para su cobertura. Bajo un esquema más alineado con la noción de capital económico [11] , el nuevo marco regulatorio propone mejoras sustanciales al tratamiento del riesgo de crédito: mayor granularidad y sensibilidad al riesgo para evaluar el riesgo, así como refinamientos en requerimientos de capital y uso de metodologías basadas en calificaciones internas (IRB, por sus siglas en inglés).

La adopción de las nuevas disposiciones regulatorias por el regulador local mexicano[12], incentivan a las instituciones de crédito a adoptar metodologías IRB para calificar su cartera crediticia. La segmentación de los clientes en distintos perfiles de crédito y la estimación de la probabilidad de incumplimiento en cada

---

[9] Reglas homogéneas que buscaban unificar las regulaciones para garantizar una mejor competencia y fijaban los requerimientos de capital como el 8% de los activos ponderados por riesgo.

[10] Adicional a los cargos de capital ya existentes para los riesgos de mercado y crédito.

[11] Capital en riesgo que una institución necesita para lograr sus objetivos de negocio suficientes para cubrir las pérdidas por los distintos riesgos que enfrenta durante un horizonte de tiempo específico (típicamente un año) y cierto nivel de confianza.

[12] La Comisión Nacional Bancaria y de Valores (CNBV) es el regulador de la banca en México.

grupo de riesgo homogéneo, son los aspectos que más impacto tienen en los requerimientos de capital y reservas que el regulador exige a las instituciones mantener para hacer frente a las pérdidas potenciales por riesgo de crédito que enfrenten sus portafolios.

Teniendo como antecedente la reciente crisis financiera, en diciembre de 2010 el BIS publicó nuevas normas regulatorias denominadas como Basilea III, las cuales han sido desarrolladas por la comunidad internacional de 27 jurisdicciones pertenecientes al BCBS, representada por 44 bancos centrales y autoridades supervisoras. Entre los aspectos más importantes que consideran las nuevas disposiciones, se encuentran iniciativas más estrictas que buscan mejorar la calidad del capital bancario, elevar el nivel exigido de capital, reducir el riesgo sistémico y establecer un nuevo marco de administración y estándares de liquidez. Asimismo, el BIS propone reforzar las exigencias de capital para riesgo de crédito proveniente de operaciones con derivados, reportos y financiamiento de valores. La implementación de tales medidas proporciona suficiente tiempo a las instituciones para una transición suave hacia el nuevo régimen.

2.2 Historia del *credit-scoring*

Mientras que los orígenes del crédito se remontan a 2000 a.C. (Asiria, Babilonia y Egipto), el credit scoring nació hace no más de 70 años. La primera aproximación para resolver el problema de identificar grupos en una población fue introducida en la estadística por Fisher (1936). Específicamente, utilizó una técnica denominada

análisis discriminante para clasificar diferentes especies de iris, la cual también empleó para discriminar entre los orígenes de cráneos, utilizando en ambos casos el tamaño físico. Durand (1941) fue el primero en reconocer que las técnicas utilizadas por Fisher (1936) también podrían aplicarse para diferenciar entre buenos y malos préstamos. Durand examinó 7200 préstamos con la técnica de análisis discriminante utilizando información sobre edad, género, estabilidad de empleo y residencia, ocupación e industria, y posesión de activos principales (cuentas bancarias, bienes raíces, seguros).

La llegada de la Segunda Guerra Mundial provocó severas pérdidas a las instituciones financieras y empresas de ventas por correo debido a la incapacidad de administrar adecuadamente su riesgo de crédito ante la escasez de analistas de crédito con suficiente experiencia. Esto llevó a la necesidad de desarrollar propiamente sistemas de scoring. La empresa Spiegel Corporation desarrolló el primer sistema de calificación crediticia (Lewis 1992). Por su parte, en 1946, en la institución Household Finance Corporation se desarrolló una "guía de calificación de crédito", la cual, no obstante su probado funcionamiento, nunca fue implantada a nivel de toda la organización (Johnson, 2004).

Para finales de la Segunda Guerra Mundial, la adopción del CS se vio frenada principalmente por la resistencia de las organizaciones al uso de las computadoras en el proceso de toma de decisiones y la ineficiencia de implementar los cálculos estadísticos y puntajes de forma manual, aunada a la dificultad de los analistas para explicar los resultados. El crecimiento del consumo y la enorme demanda del

crédito mediante productos tradicionales como la tarjeta de crédito, obligó a las instituciones a buscar alternativas de aprobación más eficientes y precisas que les permitiera administrar adecuadamente el proceso de originación para cientos de miles de nuevas solicitudes que demandaban algún tipo de crédito.

Johnson (2004) enfatiza la adopción de sistemas de juicio experto que proporcionaban cierta consistencia en el proceso de otorgamiento de crédito. La compañía Sears retomó el uso de modelos estadísticos en su negocio en la creación de sistemas de puntaje para asignar los envíos de catálogos a clientes potenciales para ventas por correo.

Entre los pioneros más conocidos del uso de sistemas de CS, se encuentra la consultoría Fair Isaac (FI) creada en San Francisco en 1956. Entre sus desarrollos iniciales se encuentra un sistema de cobranza para una tarjeta de crédito ofrecida por los hoteles Hilton. Dos años más tarde, produjeron su primera aplicación de puntajes (*scorecards*) para la empresa American Investments (Lewis, 1992).

A mediados de los 1960s las compañías de petróleo experimentaron problemas con sus operaciones de crédito principalmente por el robo de tarjetas, fraude y pérdidas por incumplimientos. Decidieron entonces adoptar metodologías más conservadoras y emplearon sistemas de CS. Las tarjetas de crédito Diners Club, American Express y Carte Blanche también implementaron este tipo de modelos en ese periodo. Por su parte, varios bancos siguieron la misma tendencia dados los altos volúmenes de solicitudes de crédito. De acuerdo con Lewis (1992), las pérdidas observadas en los portafolios fueron el factor determinante para la adopción de los

sistemas de CS. Esto permitió tomar mejores decisiones de otorgamiento de crédito, reduciendo las tasas de incumplimiento hasta en 50%.

La evolución de la regulación en los sistemas financieros terminó por afianzar la aceptación total de los sistemas de calificación crediticia en las instituciones. Por ejemplo, la aparición en EUA de los *Equal Credit Opportunity Acts* y sus reformas en 1975 y 1976, prohibieron las prácticas de discriminación en el proceso de otorgamiento de crédito a menos que las decisiones "fueran derivadas empíricamente y estadísticamente válidas".

Para los 1980s, el éxito del credit scoring en tarjetas de crédito indujo a los bancos a emplear la misma base metodológica para colocar otros productos. Al mismo tiempo, en muchas otras economías, como el Reino Unido, se dieron cambios significativos en la manera de otorgar crédito: oferta de productos a nuevos sectores, crecimiento exponencial en las tarjetas de crédito y otros créditos al consumo (Thomas et al., 2002)

Las técnicas estadísticas utilizadas en los primeros desarrollos fueron el análisis discriminante y modelos lineales, sin embargo, los desarrollos en el poder de cómputo y software estadístico de 1980s permitieron la inclusión de otras metodologías. La regresión logística, programación lineal y los árboles de decisión han sido las técnicas más utilizadas en los sistemas de calificación crediticia comerciales. Más recientemente, técnicas de inteligencia artificial como los sistemas expertos y las redes neuronales también han sido utilizadas, entre otras.

2.3 Técnicas estadísticas y de aprendizaje maquinal utilizadas en CS: Revisión de la literatura

Históricamente, el análisis discriminante y la regresión lineal han sido las técnicas más utilizadas en la construcción de sistemas de CS. Típicamente los coeficientes y puntajes numéricos de los atributos se combinan para obtener una sola contribución que se incorpora en un puntaje global. Otras técnicas utilizadas en la industria y la literatura incluyen regresión logística y otras regresiones no lineales, árboles de decisión, métodos no paramétricos como k-vecinos más cercanos, análisis de supervivencia, clasificadores bayesianos, redes neuronales, máquinas de soporte vectorial, sistemas basados en lógica difusa y modelos híbridos, entre otros.

En esta sección se realiza una revisión de la literatura relativa a distintas técnicas de clasificación que se han empleado en el contexto del CS. En el capítulo 3 se proporciona un mayor detalle sobre la formulación analítica y métodos de entrenamiento utilizados para las distintas técnicas.

2.3.1 Análisis discriminante y regresión lineal

El primer trabajo publicado sobre la aplicación del análisis discriminante lineal (LDA, por sus siglas en inglés) para construir un sistema de calificaciones se remite a Durand (1941) quien mostró que esta metodología podría producir buenas predicciones sobre el repago de los créditos. Posteriormente, otras aplicaciones relacionadas fueron las de Myers y Forgy (1963) quienes compararon los resultados

del análisis discriminante y la regresión lineal, Lane (1972), Apilado et al. (1974), Taffler y Abassi (1984), y Moses y Liao (1987).

La regresión lineal también ha sido utilizada en los modelos de calificación crediticia. Orgler (1970) utilizó regresión lineal para construir un modelo de créditos comerciales, en tanto que esta técnica también fue empleada en Orgler (1971) para construir un *scorecard* que evaluara exclusivamente el comportamiento del nivel crediticio de los clientes actuales. Otros estudios sobre el uso de modelos de regresión son Fitzpatrick (1976), Lucas (1992) y Henley (1995).

Existen otros trabajos relacionados con el estudio del comportamiento de la calidad crediticia de las empresas en función de ciertas razones financieras, como por ejemplo, Altman (1968) que construye la denominada función Z de Altman mediante el uso de LDA para pronosticar con tres años de anticipación la probabilidad de incumplimiento de las empresas. Por otro lado, Kumar y Bhattacharya (2006) comparan el desempeño entre el análisis discriminante lineal y las redes neuronales.

2.3.2 Vecinos más cercanos

El método de k-vecinos más cercanos es una aproximación no paramétrica al problema de clasificación propuesta por Fix y Hodges (1952). En el contexto del CS, primero fue aplicada por Chatterjee y Barcun (1970). Destacan varias aplicaciones posteriores, entre las que se encuentran: Henley y Hand (1996) quienes realizaron una investigación detallada de métodos de vecinos más cercanos para

datos de una compañía de ventas por correo, al tiempo que examinaron la selección de la métrica para definir la propiedad de cercanía y la selección del número de vecinos más cercanos a considerar; Paredes y Vidal (2000) evalúan la mejora en el desempeño del clasificador de vecinos más cercanos con la inclusión de una medida de disimilitud ponderada en varios conjuntos de datos, entre ellos dos correspondientes al otorgamiento de créditos; Islam et al. (2007) realiza una comparación contra el clasificador *naive Bayes* para datos de aprobaciones de tarjetas de crédito; Marinakis et al. (2008) aplican algoritmos meta-heurísticos para clasificar 1,411 empresas de un portafolio de créditos de un banco en Grecia mediante vecinos más cercanos y realizan una comparación del método contra máquinas de soporte vectorial y árboles de clasificación, entre otros. En lo general, los trabajos anteriores posicionan al clasificador de vecinos más cercanos como una alternativa competitiva frente a otros clasificadores.

A pesar de ser un método ampliamente utilizado en múltiples aplicaciones relativas a otras áreas, el clasificador de k-vecinos más cercanos no ha sido adoptado tan ampliamente en el contexto de CS, entre otras razones, debido a la percepción en la demanda computacional que requiere (Hand, 1997), aspecto que cada vez resulta un problema menor ante los avances en el cómputo actual. Otro aspecto en contra de este método tiene que ver con la imposibilidad de construir un *score* para las características particulares de cada cliente.

2.3.3 Regresión logística

El modelo logístico es una aproximación paramétrica comúnmente utilizada en la literatura para propósitos de clasificación. Barniv y McDonald (1999) reportaron 178 artículos en revistas de contabilidad y finanzas que entre 1989 y 1996 utilizaron este modelo.

El enfoque de regresión lineal para el análisis discriminante tiene la desventaja de que el rango de valores que pueden tomar los valores estimados de la variable dependiente varían sobre todos los números reales, siendo que la variable dependiente (probabilidad) se encuentra entre 0 y 1. Para solucionar este inconveniente, la regresión logística aplica la transformación $\log(p/(1-p))$, donde p es la probabilidad de que el solicitante del crédito incumpla, de manera que la variable dependiente resultante toma valores en todos los reales. Entre las ventajas de la regresión logística se encuentran (i) diseñada para trabajar con valores binarios; y (ii) los puntajes obtenidos se pueden fácilmente convertir o calibrar en estimaciones de probabilidades, dada la información disponible.

Wiginton (1980) fue uno de los primeros en publicar resultados sobre el uso de la regresión logística en los modelos de CS. A la fecha, esta técnica estadística ha sido ampliamente utilizada en las metodologías de calificación crediticia. Algunos de los trabajos en la materia son: Steenackers y Goovaerts (1989) que propone un modelo de CS para préstamos personales; Platt y Platt (1990) aplicó el análisis logístico para predecir incumplimientos con resultados interesantes en términos del desempeño de la clasificación; Laitinen (1999) pronosticó la insolvencia de 3200

empresas finlandesas utilizando 15 variables tomadas de un conjunto de 35 mediante procedimientos de selección automática bajo modelos de regresión lineales y logísticos; Moody´s (2000) aplicó la regresión logística para pronosticar el incumplimiento de 4655 empresas públicas europeas de 26 países.

Algunos estudios han encontrado un desempeño inferior en problemas de clasificación, y específicamente en sus resultados de predicción, comparado con otras técnicas. Estos hallazgos han sido corroborados, por ejemplo, por Caiazza (2004) quien comparó LDA, regresión logística, árboles de clasificación, redes neuronales y algoritmos de lógica difusa. Borra y Caiazza (2002) compararon los modelos logísticos y árboles de clasificación utilizando modelos aditivos generalizados (*bagging* y *boosting*), en tanto que Galindo y Tamayo (2000) los compararon con árboles de clasificación, redes neuronales y k-vecinos más cercanos para datos de créditos hipotecarios. Una posible explicación de estos pobres resultados es proporcionada por Alfo et al. (2005) quienes argumentan que cuando el riesgo de incumplimiento es extremadamente alto o bajo, el modelo logístico presenta fallas. Como alternativa, estos autores proponen extender el modelo bajo estructuras aleatorias que incorporen desviaciones respecto del supuesto de que los residuales provengan de un proceso de ruido blanco, mediante su denominado modelo logístico de efectos aleatorios. Tang et al. (2005) comparan los modelos logístico y de lógica difusa.

Los modelos paramétricos LDA y regresión logística han sido también aplicados en estudios realizados por bancos centrales en Austria, Francia, Alemania, Italia, Reino Unido (ver Ooghe et al., 1999).

Entre algunos de los primeros trabajos que utilizaron otros tipos de regresión no lineal se encuentra por ejemplo, Grablowsky y Talley (1981) compararon el análisis discriminante y el análisis probit[13] utilizando datos de una cadena departamental de EUA.

2.3.4 Árboles de clasificación

Los árboles de clasificación se han desarrollado en varias disciplinas, entre las que destacan las ciencias de la vida, la estadística y la inteligencia artificial. La selección de los atributos para la subdivisión en cada nodo interno es extremadamente importante en el proceso de construcción del árbol de decisión y determina en gran medida la estructura final del mismo. El primer algoritmo para generar un árbol de decisión (ID3) fue introducido por Quinlan (1979), posteriormente aparecieron el C4 de Quinlan (1986) y el C4.5 en Quinlan (1993). Quinlan utilizó el concepto ganancia de información tomando como base una medida de entropía para desarrollar estos algoritmos. Este antecedente ubica como origen de los árboles de decisión al área de aprendizaje maquinal. Sin embargo, el uso de los árboles como herramienta de clasificación y regresión mantiene una

[13] El modelo de regresión probit utiliza a la inversa de la función de distribución acumulada normal estándar como transformación que aplica sobre la probabilidad de clase para establecer la relación con los atributos o variables explicativas mediante un modelo lineal (Alpaydin, 2010).

estrecha relación con el área de la estadística. Al respecto, una de las referencias más importantes en dicho campo es Breiman et al. (1984) quienes utilizaron como criterio de selección el índice de Gini e introdujeron el término el concepto de árbol de regresión y clasificación (CART, por sus siglas en inglés). Autores como Kass (1980) y Biggs y Ville (1991) utilizaron una prueba χ^2 para realizar particiones multinivel sobre datos categóricos con su metodología denominada CHAID.

El uso de los árboles de clasificación en el contexto de CS tuvo sus inicios con Makowski (1985) y Coffman (1986) y su uso en la industria continua dada su capacidad de identificar patrones de comportamiento entre los atributos y categorías de clientes ya sea extremadamente riesgosos o con perfil de riesgo muy bajo. Anderson (2007) destaca su trasparencia y facilidad de implementación para árboles simples pero critica su poca efectividad en capacidad predictiva. Sin embargo, destaca su uso para la exploración rápida de los datos, ya sea para identificar variables predictivas clave o bien para funcionar como *benchmark* para otras metodologías.

Algunas referencias más recientes relacionadas con CS han utilizado como base de sus metodologías diversos árboles de clasificación, por ejemplo, Mues et al. (2004) utilizan diagramas de decisión (una generalización de los árboles de decisión) y explotan su uso como herramienta de descripción visual compacta de los datos. Lee et al. (2006) identificó mejores resultados con CART y regresión multivariada adaptativa por *splines* (MARS, por sus siglas en inglés) que alternativas como LDA, regresión logística, redes neuronales y máquinas de soporte vectorial para

datos de tarjeta de crédito. Zhao (2007) aplica técnicas de programación genética multi-objetivo para desarrollar árboles de decisión. Bastos (2008) implementa árboles de decisión bajo el clasificador híbrido *boosting* y muestra para datos de tarjeta de crédito que la metodología ofrece resultados con desempeño similar a preceptrón multicapa y máquinas de soporte vectorial. Por su parte, Li et al. (2010) aplica CART para analizar el comportamiento de fallas de negocio en compañías públicas listadas en la Bolsa de Shanghái (China).

2.3.5 Análisis de supervivencia

Además de ser utilizado como herramienta para pronosticar la probabilidad de incumplimiento, el análisis de supervivencia ha sido aplicado para estimar el momento en que ocurre el evento de incumplimiento de los créditos. En consecuencia, este tipo de modelos resultan útiles para estimar la rentabilidad de los clientes en los distintos productos de crédito ya que no sólo se puede analizar el riesgo de incumplimiento sino también el efecto de eventos como el prepago. Narain (1992) fue uno de los primeros en sugerir el uso de este tipo de modelos en el contexto de CS, en tanto que Banasik et al. (1999), combinó el análisis de supervivencia con regresión logística para estimar el momento de incumplimiento y el comportamiento de pago anticipado de los clientes. Otros estudios que también enfocaron sus esfuerzos en el uso de modelos de supervivencia son Hand y Kelley (2001), Stepanova y Thomas (2002), en los cuales se desarrollaron herramientas para construir *scorecards* y analizar el comportamiento de los clientes a lo largo del

tiempo[14]. Sohn et al. (2006) abordan el problema de inferencia del perfil de riesgo de los clientes no aceptados en el proceso de otorgamiento de crédito mediante un modelo de supervivencia que analiza la muestra de clientes aceptados. Sarlija et al. (2009) comparan el desempeño del análisis de supervivencia y redes neuronales en el análisis de CS para el comportamiento de conjunto de datos de créditos de un banco croata. Fantazzini y Figini (2009) proponen un método no paramétrico basado en métodos de supervivencia aleatorio para pronosticar la probabilidad de incumplimiento de empresas medianas. Los autores encuentran que el modelo mejora los resultados de clasificación en el conjunto de entrenamiento respecto del modelo logístico, en tanto que en el conjunto de prueba los resultados se invierten. Finalmente, conviene señalar que este enfoque de clasificación no será utilizado en este trabajo de investigación.

2.3.6 Redes neuronales

Las redes neuronales artificiales (NN, por sus siglas en inglés) representan una alternativa importante a los modelos estadísticos tradicionales por su capacidad como herramienta de clasificación. El caso de los modelos de credit scoring, y en general las aplicaciones en negocios, no son la excepción. De acuerdo con Vellido et al. (1999) más del 75% de la aplicaciones en este campo están relacionadas con el uso de perceptrones multicapa con propagación hacia adelante (MLP) y de retropropagación (BP).

[14] El análisis del comportamiento de los clientes a lo largo del tiempo también es conocido como calificación de comportamiento (*behavioral-scoring*).

Jensen (1992) aplica MLP para pronosticar el comportamiento de pago de aplicaciones de tarjetas de crédito y lo comparó contra un modelo de CS comercial; West (2000) comparó el desempeño de cinco técnicas de NN en los conjuntos de créditos Alemán y Australiano[15]; por su parte, West et al. (2005) incrementan la capacidad de los modelos anteriores mediante las estrategias de validación cruzada, *bagging* y *boosting*.

No obstante el buen desempeño de las tasas de clasificación observadas en las técnicas de NN, se les critica por la ausencia de variables explicativas sobre la clasificación de los créditos y su poca transparencia explicativa (Baesens et al., 2003a). Algunos de los métodos propuestos en la literatura para solucionar este problema son la combinación con otros métodos, extracción de reglas de las redes entrenadas y métodos de agrupamiento. Sobre esta última estrategia, Huysmans et al. (2006) explotó la capacidad de visualización y análisis exploratorio de los mapas auto organizados (SOM) para el caso de CS y determinó un mejor desempeño cuando se utilizaba para elevar el poder predictivo de cada neurona en un sistema de NN. En lo relativo a la extracción de reglas, Baesens et al. (2003a) aplicó tres metodologías para la extracción de reglas que simulen el proceso de decisión de NN entrenadas (a saber: NeuroRule, Trepan y Nefclass[16]) para clasificar clientes buenos

[15] Las técnicas utilizadas por este autor fueron: MLP, mezcla de sistemas expertos, función radial, vector de cuantización y resonancia adaptativa difusa.

[16] El algoritmo de descomposición NeuroRule extrae reglas proposicionales minimizando las suma de cuadrados de los errores en las predicciones (Setiono y Liu, 1996). Por su parte, el algoritmo Trepan, introducido por Craven y Shavlik (1996), induce árboles de decisión a partir de NNs entrenadas. Por su parte, Nefclass extrae reglas difusas interpretables (Nauck et al., 1997).

y malos de tres distintos conjuntos de créditos, realizando una comparación contra el algoritmo C4.5 y regresión logística. Asimismo, Arns et al. (2006) utilizaron la técnica de NeuroRule para extraer reglas de NNs multicapa entrenadas para modelar el perfil de riesgo de portafolio de créditos a pequeñas y medianas empresas otorgados por un banco brasileño.

2.3.7 Máquinas de soporte vectorial

Las máquinas de soporte vectorial (SVM) han sido utilizadas para problemas de clasificación y regresión debido a su buen desempeño en diversas aplicaciones. Recientemente, varios estudios han utilizado SVM en credit scoring con resultados alentadores. El problema de la predicción del incumplimiento de empresas también ha sido abordado aplicando NN y SVM en Fan y Palaniswami (2000), Atiya (2001), Shin et al. (2005) y Min et al. (2005). Li et al. (2006) desarrolló un modelo de CS para aplicaciones de créditos al consumo identificando mejorías en la generalización respecto a modelos de NN. Para solucionar el problema sobre la interpretación de los resultados de clasificación de SVM, Martens et al. (2007) introdujo métodos de extracción de reglas para SVM que pueden generar modelos de credit scoring con reglas interpretables y con poca pérdida en la exactitud.

En un estudio comparativo de 17 clasificadores, Baesens et al. (2003b) analizaron 8 conjuntos de datos de créditos, entre los que se encuentran algunos portafolios de instituciones financieras del Reino Unido, Bélgica, Holanda y Luxemburgo. Los métodos utilizados fueron: regresión lineal y cuadrática, regresión logística,

programación lineal, cuatro variantes de SVM, cuatro clasificadores basados en árboles de decisión, dos variantes de vecinos más cercanos, NN, *naive Bayes*, y árboles bayesianos aumentados bajo *naive Bayes*. Los resultados obtenidos por los autores indicaron que los métodos con mejor desempeño fueron SVM con base radial y la versión por mínimos cuadrados, así como también los clasificadores basados en NN.

2.3.8 Clasificadores Bayesianos

El clasificador más simple de los clasificadores bayesianos es el denominado *naive Bayes*, cuya efectividad en distintos contextos y áreas de aplicación lo hacen un clasificador *benchmark* obligado en los distintos estudios de CS realizados en la literatura (Alpaydin, 2010). Por su parte, las redes bayesianas (BN) son un modelo probabilista gráfico que representa a un conjunto de variables aleatorias cuyas dependencias condicionales se establecen vía una gráfica acíclica dirigida. El texto seminal de Pearl (1988) estableció a las BN como un campo de estudio al consolidar sus propiedades. Este tipo de redes representan un conjunto de clasificadores que modelan las relaciones estructurales de las variables. Davis et al. (1992) investigaron mecanismos bajos sistemas expertos bayesianos podrían ayudar al problema de clasificación de tarjetas de crédito de Bank of Scotland y comparó la aproximación con NN. Baesens et al. (2002) evalúa la aplicabilidad de varios tipos de BN en CS, entre ellos, el clasificador *naive Bayes*, árboles aumentados bajo *naive Bayes*, y una BN irrestricta con entrenamiento Monte Carlo via Cadenas de

Markov (MCMC, por sus siglas en inglés). En un ejercicio empírico sobre el conjunto de datos Alemán, este último clasificador tuvo el mejor desempeño comparado con el resto y el algoritmo C4.5. Por su parte, Li y Guo (2006) determinan capacidades similares entre tres modelos de clasificación basados en NB y cinco tipo de NN, sobre dos conjuntos de datos de créditos reales. Más recientemente, Antonakis y Sfakianakis (2009) comparan modelos de credit scoring construidos sobre la NB contra LDA, regresiones logísticas, vecinos más cercanos, árboles de clasificación y NN. Los autores utilizan dos conjuntos de datos, uno con créditos de un banco griego y el segundo formado por el conjunto de datos Australiano, encontrando un poder predictivo inferior del clasificador NB respecto a los otros métodos.

2.3.9 Sistemas basados en lógica difusa

Un sistema difuso es un conjunto de reglas difusas IF-THEN que convierte las entradas en salidas. El motor de inferencia difuso combina reglas mediante un mapeo de los conjuntos difusos en el espacio de entrada X al conjuntos difusos en el espacio de salida Y basado en principios de lógica difusa (Bethold, 1999). Malhotra y Malhotra (2002) utilizaron un algoritmo de lógica difusa llamado ANFIS para construir un modelo de credit scoring el cual mostró un mejor desempeño que LDA, con mayor flexibilidad, tolerancia a información incompleta y capaz de modelar relaciones no lineales complejas, sin embargo, una deficiencia se relacionaba con su alto costo computacional. Piramuthu (1999) aplica un sistema difuso para

desarrollar un modelo de credit scoring sobre tres diferentes conjuntos de datos el cual compara con un modelo de NN, reportando un mejor desempeño de este último en términos de clasificación tanto en el conjunto de entrenamiento como en el de prueba con la diferencia de que el modelo de lógica difusa sí tiene interpretación para el usuario.

Wang et al. (2005) aplicó SVM bajo lógica difusa bajo el supuesto de que los clientes no pueden ser absolutamente buenos o absolutamente malos. Su estudio contempla también el uso de regresión lineal, regresión logística y redes de retro-propagación.

2.3.10 Programación genética

La programación genética (GP), un tipo de técnica evolutiva, ha sido aplicada también para propósitos de calificación crediticia. Por ejemplo, Ong et al. (2005) concluyen que los modelos de credit scoring basados en GP ofrecen mejores resultados que los enfoques que utilizan NN, árboles de decisión, conjuntos rugosos y regresión logística. En otro estudio relativo a credit scoring, Huang et al. (2006) aplican un modelo de GP en dos etapas que incorpora las ventajas de las reglas IF-THEN y la función discriminante, encontrando evidencia de un mejor desempeño frente a CART, C4.5, conjuntos rugosos y regresión logística. En ambos estudios los conjuntos de datos utilizados son los conjuntos Alemán y Australiano.

2.3.11 Modelos híbridos

Con referencia a la mezcla de muchas de las técnicas estadísticas y de aprendizaje maquinal utilizadas en la literatura, y la mayoría de las veces poco explotadas en la industria, varios autores han propuesto modelos híbridos para atacar el problema de CS. Por ejemplo, Zhu et al. (2001) estudió las condiciones bajo las cuales se pueden combinar distintos clasificadores para construir un modelo combinado que mejore los resultados individuales. Entre las técnicas comúnmente utilizadas para realizar clasificadores múltiples (*ensemble*), se pueden citar las denominadas *bagging, boosting* y *stacking*. Al respecto, Wang et al. (2011) realiza una evaluación comparativa del desempeño entre las tres técnicas citadas. Algunos trabajos relacionados con modelos híbridos aplicados en calificación crediticia son Lee y Chen (2002), Lee et al. (2002), Hsieh (2005) y más recientemente Li (2009). Por mencionar algunos otros, Sexton et al. (2006) utilizó un método basado en algoritmos genéticos llamado algoritmo de optimización simultánea de NN para datos sobre aprobación de tarjetas de crédito. Huang et al. (2007) desarrolló un modelo híbrido basado en SVM y algoritmos genéticos capaz de implementar tareas de selección y optimización de parámetros, el cual presenta resultados similares a los de clasificadores basados en redes neuronales, árboles de decisión y programación genética, y fue aplicado a dos conjuntos de datos. Martens et al. (2008) extraen reglas de decisión para construir un sistema de calificación interno utilizando principios de optimización hormiga empleando el algoritmo AntMiner+ sobre un conjunto de créditos a empresas medianas. Mileris y Boguslauskas (2010)

analizan el impacto que tiene la técnica de reducción de datos con análisis de factores (ANOVA y K-S) sobre la predicción de incumplimientos de compañías en Lituania bajo LDA, regresión logística y MLP. Ghodselahi (2011) propone un modelo híbrido que combina técnicas de agrupamiento y clasificación sobre el conjunto de créditos alemanes de UCI, específicamente, utiliza 10 clasificadores SVM como miembros de un clasificador *ensemble* que combina varias hipótesis construidas sobre el conjunto de entrenamiento.

En la siguiente figura se incluyen varias de las técnicas de clasificación mencionadas en esta sección, así como los primeros trabajos en los que fueron aplicadas en el contexto de CS.

	Técnica	Trabajos iniciales
Estadística	Análisis discriminante	Durand (1941)
	Vecinos más cercanos	Chatterjee y Barcun (1970)
	Regresión logística	Wiginton (1980)
	Árboles de clasificación	Makowski (1985)
	Análisis de supervivencia	Narain (1992)
Reconocimiento Patrones	Clasificadores bayesianos	Davis et al. (1992)
	Redes neuronales	Jensen (1992)
	Máquinas de soporte vectorial	Fan y Palaniswami (2000)
	Lógica difusa	Piramuthu (1999)
	Programación genética	Ong et al. (2005)
	Modelos híbridos	Zhu et al. (2001), Lee et al. (2002)

Tabla 1. Técnicas de clasificación utilizadas en CS y primeros trabajos

Es importante mencionar que existe una delgada línea entre las áreas de estadística de reconocimiento de patrones con respecto a algunos de los clasificadores incluidos en la Tabla 1. La taxonomía utilizada no se debe considerar como absoluta y definitiva. Por ello, se incluyeron a la izquierda de la tabla los encabezados la parte central en colores difuminados para indicar los casos en que los clasificadores considerados han tenido desarrollos paralelos en ambas ramas, y que diversos autores, dependiendo del enfoque y uso que les dan a los mismos, los clasifican indistintamente como pertenecientes a alguna de las dos áreas.

En este capítulo se realizó una revisión de aspectos fundamentales relacionados con CS, desde la definición básica de crédito y calificación crediticia, hasta aspectos de tipo regulatorio en torno al tema. Por otra parte, se presentó una revisión de la literatura sobre las distintas técnicas que han sido utilizadas para la construcción de modelos de CS, tanto en la industria como en la literatura.

En el siguiente capítulo se describen los detalles metodológicos subyacentes a los clasificadores que se utilizarán en este trabajo de investigación para la construcción de modelos de CS. Es conveniente mencionar que tanto los sistemas basados en lógica difusa como los de programación genética no serán utilizados en el ejercicio empírico de este documento.

Capítulo 3

Modelos estadísticos y de reconocimiento de patrones

En este capítulo se realiza una revisión sobre los fundamentos de algunos de los principales modelos estadísticos y de reconocimiento de patrones que se han utilizado en la literatura e industria para el problema de clasificación en el contexto del *credit-scoring* (CS).

Para ello, se considera un conjunto $D = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{N}$ de N observaciones (ejemplos), cuyos n atributos están contenidos en la componente n-dimensional $\boldsymbol{x_i} \in X$ y las etiquetas de clases binarias $y_i \in C \equiv \{0,1\}$.

El proceso de entrenamiento de un clasificador $f: X \rightarrow C$ consiste en configurar los parámetros que lo caracterizan, mediante alguna técnica de estimación o aprendizaje sobre la base de cierto conjunto de entrenamiento D. El resultado del

proceso de estimación generalmente es un mapeo[17] $\hat{f}$ que busca aproximar la regla de correspondencia $\boldsymbol{x_i} \mapsto y_i, i = 1, \ldots, N$.

3.1 Regresión logística y análisis discriminante lineal

Supóngase que se tienen etiquetas de clases binarias y_i. La estrategia de la regresión logística para clasificación (LOG) consiste en estimar la probabilidad $P(y = 1|x)$ mediante la función sigmoide

$$P(y = 1|\boldsymbol{x}) = \frac{1}{1+\exp\left(-(w_0+\boldsymbol{w}^{\mathrm{T}}\mathrm{x})\right)} \qquad (1)$$

donde $\boldsymbol{x} \in \mathbb{R}^n$ es un vector de entrada n-dimensional, $\boldsymbol{w}$ es el vector de parámetros y el escalar w_0 es el intercepto (Baesens, 2003b). Los parámetros w_0 y $\boldsymbol{w}$ son típicamente estimados utilizando el procedimiento de máxima verosimilitud. Si $l(w_0, \boldsymbol{w})$ denota la función de verosimilitud, se puede considerar la maximización de la función penalizada[18] $l(w_0, \boldsymbol{w}, R) = l(w_0, \boldsymbol{w}) - R\|(w_0, \boldsymbol{w})\|^2$. El uso del parámetro R denominado *ridge* estabiliza los casos degenerados y permite reducir la el sobre-ajuste penalizando valores grandes de los coeficientes de la regresión (ver Cessie y van Houwelingen, 1992).

---

[17] La regla de correspondencia definida por f contempla el uso de atributos tanto numéricos como no numéricos contenidos en el conjunto X.

[18] El operador $\|\cdot\|$ representa la norma euclidiana, la cual se define como $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}$ para todo vector $\boldsymbol{x} \in \mathbb{R}^n$.

Por su parte, el análisis discriminante asigna una observación $\boldsymbol{x}$ a la clase $y_i \in \{0,1\}$, teniendo la mayor probabilidad posterior $p(y|x)$. Por el teorema de Bayes se puede realizar el cómputo de la probabilidad posterior como

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} \tag{2}$$

Si se supone que las distribuciones condicionales $p(\mathbf{x}|y)$ son Gaussianas multivariadas se puede verificar que

$$p(\mathbf{x}|y=1)\frac{1}{(2\pi)^{n/2}|\Sigma_1|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_1})^{\mathrm{T}}\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu_1})\right\} \tag{3}$$

donde $\boldsymbol{\mu_1}$ es el vector de medias de la clase 1 y Σ_1 denota su matriz de varianza-covarianza. Entonces bajo el criterio de clasificación discriminante, se asigna la clase 1 si

$$(\mathbf{x}-\boldsymbol{\mu_1})^{\mathrm{T}}\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu_1}) - (\mathbf{x}-\boldsymbol{\mu_0})^{\mathrm{T}}\Sigma_0^{-1}(\mathbf{x}-\boldsymbol{\mu_0}) <$$

$$2log\big(P(y=1)\big) - 2log\big(P(y=0)\big) + log|\Sigma_0| - log|\Sigma_1|, \tag{4}$$

y la clase 0 en otro caso. La presencia de términos $\mathbf{x}^{\mathrm{T}}\Sigma_1^{-1}\mathbf{x}$ y $\mathbf{x}^{\mathrm{T}}\Sigma_0^{-1}\mathbf{x}$ indican que el límite de decisión es cuadrático en x y por lo tanto, esta técnica de clasificación es denominada *análisis discriminante cuadrático (CDA)*. En el caso en que $\Sigma_0 = \Sigma_1 = \Sigma$, los términos cuadráticos $\mathbf{x}^{\mathrm{T}}\Sigma_1^{-1}\mathbf{x}$ y $-\mathbf{x}^{\mathrm{T}}\Sigma_1^{-1}\mathbf{x}$ se cancelan y la regla de clasificación se vuelve lineal en $\boldsymbol{x}$. A la técnica de clasificación obtenida comúnmente se le refiere como *análisis discriminante lineal (LDA)*.

3.2 Máquinas de soporte vectorial (SVM)

De acuerdo con la formulación original de Vapnik (1998), si se emplean etiquetas de clases binarias $y_i \in \{-1, +1\}$, el clasificador SVM, debe satisfacer las siguientes condiciones:

$$\begin{cases} \boldsymbol{w}^{\mathrm{T}}\varphi(\mathbf{x}_i) + \mathrm{b} \geq +1 & si \;\; y_i = +1 \\ \boldsymbol{w}^{\mathrm{T}}\varphi(\mathbf{x}_i) + \mathrm{b} \leq -1 & si \;\; y_i = -1 \end{cases} \tag{7}$$

El sistema (7) es equivalente a

$$y_i[\boldsymbol{w}^{\mathrm{T}}\varphi(\mathbf{x}_i) + \mathrm{b}] \geq 1, \qquad i = 1, \ldots, N \tag{8}$$

La función no lineal $\varphi(\cdot)$ mapea el espacio de atributos en un espacio de dimensión mayor (posiblemente infinita). En este espacio de dimensión superior, las desigualdades presentadas previamente, básicamente construyen un híper-plano $\boldsymbol{w}^{\mathrm{T}}\varphi(\mathbf{x}_i) + \mathrm{b} = 0$ que discrimina entre las dos clases. En el espacio primal de los pesos, el clasificador adquiere la siguiente forma:

$$y(x) = sign[\boldsymbol{w}^{\mathrm{T}}\varphi(\mathbf{x}_i) + \mathrm{b}], \tag{9}$$

Esencialmente, los clasificadores SVM utilizan modelos lineales para establecer regiones que separen clases de manera no lineal. La estrategia que utilizan consiste en transformar el espacio original de los atributos de entrada mediante el mapeo no lineal φ, de manera que en el nuevo espacio las clases queden separadas por un hiperplano. En consecuencia, las regiones de separación en el espacio original son no lineales.

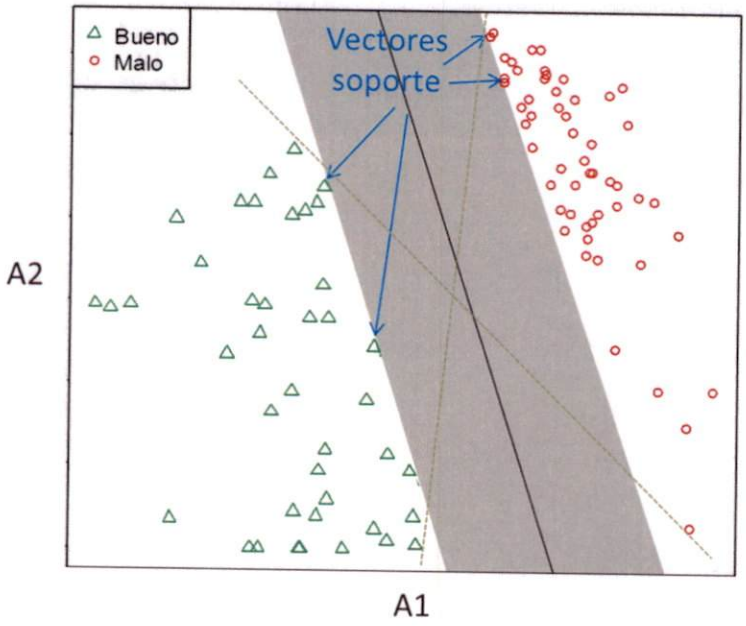

Figura 1. Conjunto linealmente separable bajo SVM (dos atributos A1 y A2; dos clases)

En la Figura 1 se presenta un ejemplo de un conjunto linealmente separable en el contexto de CS respecto a dos atributos dados A1 y A2.

No obstante la existencia de una regla de separación dada por la ecuación (9), la evaluación no se realiza de esta manera. En cambio, define el problema de optimización convexa

$$\min_{w,b,\xi} \mathcal{J}(\mathbf{w}, b, \xi) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i \qquad (10)$$

sujeto a

$$\begin{cases} y_i[\mathbf{w}^T\varphi(\mathbf{x}_i) + \mathbf{b}] \geq 1, & i = 1, \ldots, N \\ \xi_i \geq 0 & i = 1, \ldots, N \end{cases} \qquad (11)$$

Las variables de holgura ξ_i se utilizan para permitir malas clasificaciones en el conjunto de desigualdades (por ejemplo, debido a distribuciones traslapadas). La

primera parte de la función objetivo busca maximizar el margen entre ambas clases en el nuevo espacio, mientras que la segunda parte minimiza el error incurrido por malas calificaciones. La constante $C > 0$ se considera como un parámetro de refinamiento en el algoritmo, también llamado parámetro de costo. Hay que notar que esta formulación esta cercanamente relacionada con un problema de programación lineal. Entre las diferencias más importantes se encuentra la introducción de un término de regularización $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$ grande en la función objetivo, asimismo, considera un margen para separar las clases y permite cotas de decisión no lineales mediante el mapeo $\varphi(\cdot)$.

El lagrangiano del problema de optimización (10) y (11) con restricciones está dada por

$$\mathcal{L}(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{v}) = \mathcal{J}(\mathbf{w}, b, \xi) - \sum_{i=1}^{N} \alpha_i \{y_i[\boldsymbol{w}^{\mathrm{T}}\varphi(\mathbf{x}_i) + \mathrm{b}] - 1 + \xi_i\} - \sum_{i=1}^{N} v_i \xi_i \quad (12)$$

La solución al problema de optimización presentado arriba, está dado por el punto silla del lagrangiano, el cual se obtiene al minimizar (12) con respecto de $\mathbf{w}, b, \xi$ y maximizándolo con respecto a $\boldsymbol{\alpha}\ y\ \boldsymbol{v}$. Para resolver el problema

$$\mathbf{max}_{\boldsymbol{\alpha},\boldsymbol{v}}\ \mathbf{min}_{\mathbf{w},b,\xi}\ \mathcal{L}(\mathbf{w}, b, \xi, \boldsymbol{\alpha}, \boldsymbol{v}) \quad (13)$$

se obtiene:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \quad \rightarrow \quad w = \sum_{i=1}^{N} \alpha_i y_i \varphi(x_i) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \quad \rightarrow \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \\ \frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \quad \rightarrow \quad 0 \leq \alpha_i \leq C, i = 1, \ldots, N \end{cases} \tag{14}$$

Al sustituir la primera expresión en $y(x) = sign[\mathbf{w}^{\mathrm{T}}\varphi(\mathbf{x_i}) + \mathrm{b}]$, el clasificador resultante quedaría dado por

$$y(x) = sign[\sum_{i=1}^{N} \alpha_i y_i \, K(\boldsymbol{x}_i, \mathbf{x}) + \mathrm{b}] \tag{15}$$

donde $K(\mathbf{x_i}, \mathbf{x}) = \varphi(\mathbf{x}_i)^{\mathrm{T}}\varphi(\mathbf{x})$ constituye un núcleo definido positivo que satisface el teorema de Mercer [19]. Los multiplicadores de Lagrange α_i son entonces determinados mediante el siguiente problema de optimización (problema dual):

$$\max_{\alpha_i} -\frac{1}{2}\sum_{i,j=1}^{N} y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)\, \alpha_i \alpha_j + \sum_{i=1}^{N} \alpha_i \tag{16}$$

sujeto a

$$\begin{cases} \sum_{i=1}^{N} \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, N \end{cases} \tag{17}$$

La construcción del clasificador SVM ahora se reduce a resolver un problema de programación cuadrática convexa en α_i. Como primer aspecto, conviene notar que no será necesario calcular $\boldsymbol{w}$ o $\varphi(x_i)$ para determinar la superficie de decisión, por ello, no será necesaria la construcción de un mapeo no lineal $\varphi(\mathbf{x})$. Como alternativa, se emplea la función núcleo K. Entre las formas funcionales más

[19] En general, se dice que toda función $K(u, v)$ define un kernel si satisface el teorema de Mercer, es decir, $\int_{u,v} K(u, v)g(u)g(v)dudv > 0$ para toda función $g(\cdot)$ cuadrado integrable (ver Mercer, 1909). El Teorema de Mercer permite interpretar las funciones kernel como el producto punto en un espacio característico.

comunes para K se tienen la función de base radial (RBF) $K(\mathbf{x_i}, \mathbf{x}) = \exp\left\{-\gamma\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right\}$ y el núcleo lineal $K(\mathbf{x_i}, \mathbf{x}) = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$. Además de los clasificadores SVM estándar, en la literatura se han propuesto los clasificadores LS-SVM, los cuales son una versión modificada de los SVMs sugeridos por Suykens y Vandewalle (1999) y Suykens et al. (2002). LS-SVM utiliza una función de costos de mínimos cuadrados y reemplazan las restricciones de desigualdad con restricciones de igualdad para resolver el problema

$$\min_{w,b,\xi} \mathcal{J}(\mathbf{w}, b, \xi) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \qquad (18)$$

sujeto a las restricciones de igualdad:

$$y_i[\boldsymbol{w}^{\mathrm{T}}\varphi(\mathbf{x_i}) + \mathrm{b}] = 1 - e_i, \qquad i = 1, \ldots, N \qquad (19)$$

Siguiendo la misma estrategia de solución que con el clasificador SVM estándar, se puede verificar que el clasificador LS-SVM se puede obtener como la solución al sistema lineal de ecuaciones:

$$\begin{pmatrix} 0 & | & \boldsymbol{y}^{\mathrm{T}} \\ - & | & - \\ \boldsymbol{y} & | & \Omega + \gamma^{-1}\mathbf{I} \end{pmatrix}\begin{pmatrix} b \\ - \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} 0 \\ - \\ \mathbf{1} \end{pmatrix} \qquad (20)$$

donde $\mathbf{y} = [y_1; \ldots; y_N], \mathbf{1} = [1; \ldots; 1], \mathbf{e} = [e_1; \ldots; e_N]$, $\boldsymbol{\alpha} = [\alpha_1; \ldots; \alpha_N]$. Entonces el teorema de Mercer se puede aplicar directamente a la matriz Ω, definida por $\boldsymbol{\Omega}_{ij} = y_i y_j \varphi(\mathbf{x_i})^{\mathrm{T}}\varphi(\mathbf{x_j}) = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$.

3.3 Redes neuronales

Las redes neuronales (NN) son representaciones matemáticas inspiradas en el funcionamiento del cerebro humano. Muchos tipos de redes neuronales han sido sugeridas en la literatura (Witten et al., 2011). Entre los ejemplos de NN más populares se encuentra el denominado perceptrón multicapa (MLP).

Un MLP está compuesto típicamente por una capa de entrada, una o varias capas ocultas y una capa de salida, cada una consistiendo de cierto número de neuronas. Cada neurona procesa sus entradas y genera un valor de salida que es transmitido a las neuronas en la capa subsecuente[20].

En la siguiente figura se muestra la configuración típica en el contexto de CS, para un MLP con una capa oculta y una capa de salida formada por una sola neurona.

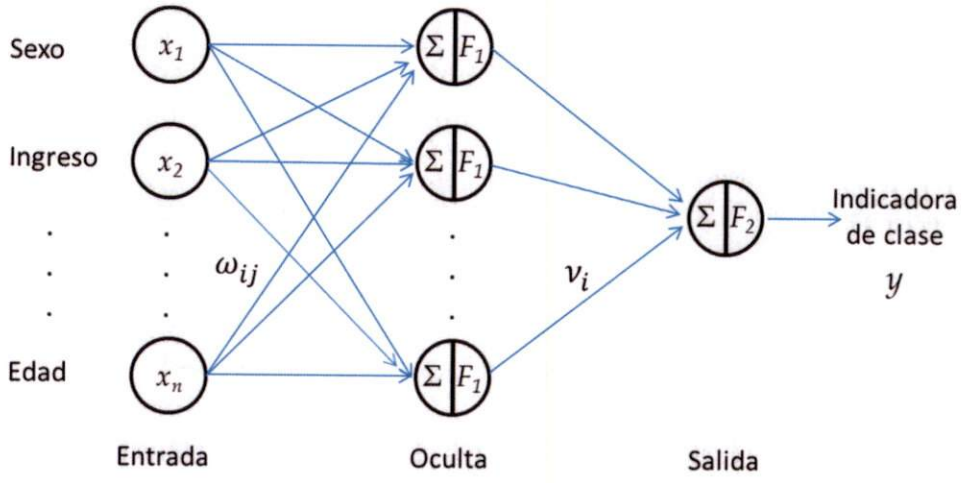

Figura 2. Perceptrón multicapa con una capa oculta

[20] Dado que en este caso las entradas alimentan directamente a las salidas vía los pesos que unen a las unidades, el MLP es considerado uno de los ejemplos más simples de las NN de tipo *feed-forward*.

En la capa de entrada se incorporan n atributos (como los que usualmente son requeridos en las solicitudes de créditos personales).

La salida h_i para cada neurona oculta i se obtiene de ponderar cada una de sus entradas (atributos) más un término de tendencia o sesgo $b_i^{(1)}$ como se muestra a continuación:

$$h_i = F_1\left(b_i^{(1)} + \sum_{j=1}^{n} \omega_{ij} x_j\right) \tag{21}$$

donde ω_{ij} denota el conector de entrada de pesos j hacia la unidad oculta i. Por su parte, como resultado de la capa de salida se construye la indicadora de clase y de la siguiente manera:

$$y = F_2\left(b^{(2)} + \sum_{i=1}^{n_h} \upsilon_i h_i\right) \tag{22}$$

donde n_h es el número de neuronas ocultas, υ_i representa el peso conector de la neurona oculta i hacia la neurona de salida y $b^{(2)}$ el término de sesgo correspondiente. Las entradas de sesgo juegan un rol similar al término de intercepto en un modelo de regresión lineal clásico.

Las funciones F_1 y F_2 que aparecen en las ecuaciones (21) y (22) se denominan funciones de transición o transferencia, las cuales permiten modelar relaciones no lineales entre los datos. Entre las funciones de transferencia más comúnmente

utilizadas se encuentran: la función escalón[21] $F(x) = 1_A(x)$, la logística $F(x) = \frac{1}{1+e^{-x}}$, la tangente hiperbólica $F(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ y la función lineal $F(x) = x$.

En el caso de un problema de clasificación binaria, es conveniente utilizar la función de transferencia logística en la capa de salida (F_2), ya que esta salida se limita a un valor entre el rango [0,1][22]. Esto permite que la salida y de la MLP sea interpretada por medio de probabilidad condicional. Finalmente, una función umbral es típicamente aplicada para mapear la red de salida y a una etiqueta de clasificación (por ejemplo, bueno/malo, en el contexto de CS).

Nótese que múltiples capas ocultas pueden ser utilizadas pero diversos resultados teóricos han demostrado que las NN de tipo *feed-forward* con una sola capa oculta son capaces de aproximar de manera universal cualquier función continua para cualquier grado de precisión deseado en un intervalo compacto (Cybenko, 1989).

Los pesos ω_{ij} y υ_i, así como los coeficientes de sesgo $b_i^{(1)}$ y $b^{(2)}$, son parámetros cruciales de la red y necesitan ser estimados durante un proceso de aprendizaje. Al respecto muchos algoritmos han sido sugeridos en la literatura, entre los que destaca el algoritmo de retro-propagación. Como cualquier otro esquema de clasificación, los MLP entrenados con retro-propagación pueden sufrir de un excesivo sobre-ajuste en el conjunto de entrenamiento, específicamente si la red es mucho más grande que la estructura necesaria para describir el problema en

[21] El conjunto A corresponde a cierto intervalo en donde la función toma el valor 1 y cero fuera del conjunto. Esta función también es conocida como función indicadora.

[22] Básicamente, cualquier función de distribución puede ser utilizada como función de transferencia para la capa de salida.

estudio. Entre las estrategias sugeridas para aminorar este efecto se encuentran: evaluar la tasa de error de clasificación para cierto subconjunto de datos y detener anticipadamente (*early stopping*) el proceso de estimación cuando ésta empiece a disminuir a partir de cierto umbral; el uso de un factor de decaimiento en el que la función de error se penaliza por la suma de cuadrados de los pesos estimados en la red, con la intención de limitar la influencia de conexiones irrelevantes entre neuronas. Adicionalmente, es común utilizar un parámetro de tasa de aprendizaje L, el cual corresponde a la proporción de pesos que son actualizados en cada iteración. El lector puede consultar Witten et al. (2011) para profundizar con mayor detalle en las estrategias de calibración de los parámetros de NNs.

Una desventaja importante del MLP es que contiene unidades internas que esencialmente son poco transparentes. No obstante que existen diversas técnicas para extraer reglas de decisión a partir de redes neuronales entrenadas, no es claro si éstas ofrecen alguna ventaja con respecto a clasificadores estándar que inducen conjuntos de reglas directamente de los datos y cuyo entrenamiento es más rápido.

3.4 Clasificadores Bayesianos

Un clasificador simple que en la práctica suele desempeñarse de manera sorprendentemente aceptable en distintos contextos es el clasificador *naive Bayes*. Este clasificador básicamente aprende de las probabilidades condicionales de las clases $p(x_i|y)$de cada entrada x_i, $i = 1, \ldots, n$ dada una etiqueta de clase y. Cada

ejemplo o instancia se clasifica utilizando la regla de Bayes para calcular la probabilidad posterior de cada clase y dado el vector $\mathbf{x}$ de los valores observados de sus atributos (Baesens et al., 2003b):

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}$$

Estos clasificadores trabajan bajo el supuesto de que los atributos son condicionalmente independientes dada la etiqueta de clase, por lo tanto,

$$p(\mathbf{x}|y) = \prod_{i=1}^{n} p(x_i|y) \qquad (23)$$

La ecuación (23) constituye una simplificación en la estructura de dependencia de los atributos, la cual facilita enormemente los cálculos. Las probabilidades $p(x_i|y)$ son estimadas usando los conteos de frecuencias para los atributos discretos, y un método bajo la distribución normal o una densidad (kernel) de suavizamiento para los atributos continuos. Los clasificadores obtenidos bajo el supuesto de la ecuación (23) se denominan clasificadores de Bayes simples (*naive Bayes*).

Friedman et al. (1997) introdujeron clasificadores de Bayes bajo un contexto de árbol aumentado (TAN) como una extensión de los clasificadores de Bayes simples. Los TANs dispensan el supuesto de independencia al permitir dependencias con estructura arbórea entre los atributos. Una dependencia desde x_i hasta x_j implica que el impacto de x_i sobre la variable de clase también depende del valor de x_j. Estos clasificadores pertenecen a un grupo más grande denominado redes bayesianas.

3.4.1 Clasificadores basados en Redes Bayesianas

Las redes bayesianas (BN) pertenecen a la familia de modelos gráficos probabilísticos, los cuales son utilizados para representar relaciones entre variables bajo contextos específicos (Witten et al., 2011). Específicamente, las BN corresponden a una estructura conocida como gráfica acíclica dirigida (DAG), cuya estructura es definida por dos conjuntos: el conjunto de nodos (vértices) y el conjunto de arcos dirigidos. Los nodos representan variables aleatorias (v.a.) las cuales son representadas por círculos y los arcos dirigidos que representan la dependencia directa entre las v.a. son dibujados como flechas entre los nodos. Cuando el arco es dibujado del nodo X_i (padre) al nodo X_j (hijo), la estructura indica que la v.a. X_j depende de X_i. La estructura acíclica de la gráfica garantiza que ningún nodo puede ser su propio antecesor o su propio sucesor, de manera que la probabilidad condicional de todas la variables puede escribirse de forma compacta. No obstante que las flechas que gráficamente unen nodos representan una conexión causal entre las variables, el proceso de razonamiento de las BN puede operar propagando información en cualquier dirección. La independencia condicional de cada variable con todos aquellos nodos que no sean sus descendientes, dado el estado de sus padres, reduce significativamente el número de parámetros necesarios para caracterizar la distribución conjunta de las variables, al tiempo de proveer una manera eficiente de calcular probabilidades posteriores.

Adicional a la estructura DAG, para caracterizar completamente a una BN, se requiere especificar los parámetros del modelo de manera tal que se cumple la propiedad de Markov y en consecuencia la probabilidad condicional en cada nodo depende solamente de sus padres. Para el caso de v.a. discretas, dicha probabilidad condicional se representa mediante una tabla que enlista, para cada combinación de los valores de sus padres, la probabilidad (local) de que un nodo hijo tome sus valores posibles.

De acuerdo con Friedman et al. (1997), una red Bayesiana B es una gráfica acíclica que representa probabilidades condicionales sobre un conjunto de v.a. $U = \{Z_1, Z_2, \ldots, Z_m\}$. La red queda definida por el par $B = \langle G, \Theta \rangle$, donde G es la DAG cuyos nodos son los elementos en U y la estructura de sus arcos representan la dependencias directas entre las v.a. Z_i. Por su parte, Θ denota al conjunto de parámetros de la red formado por $\theta_{z_i|\pi_i} = P_B(z_i|\pi_i)$ para cada realización z_i de Z_i condicionada sobre π_i, el conjunto de padres de Z_i en G. Si Z_i no tiene padres, se dice que su distribución de probabilidades locales es incondicional, y en cualquier otro caso, se denomina condicional. Entonces

$$P_B(U) = \prod_{i=1}^{m} \theta_{Z_i|\pi_i} \tag{24}$$

La ecuación (24) permite calcular la distribución conjunta de U de manera más compacta que si la comparamos con la regla de la cadena inducida por la definición de probabilidad condicional

$$P_B(U) = \prod_{i=1}^{m} P_B(Z_i|Z_{i+1}, \ldots, Z_m), \tag{25}$$

donde el último término del producto correspondiente al índice $i = m$ es $P_B(Z_m)$.

La factorización (24) ofrece ventajas para la inferencia, aprendizaje (estimación de parámetros) y también desde el punto de vista computacional. De esta forma, cualquier probabilidad conjunta o condicional de interés sobre cualesquier subconjunto de v.a. no observables se puede calcular sumando/integrando a (24) sobre todos los valores posibles (observados) del complemento del subconjunto de interés.

Siguiendo la notación introducida al inicio del capítulo, se dispone de un conjunto de datos $D = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{N}$, donde cada y_i y $\boldsymbol{x_i}$ representan realizaciones de la v.a. de clase Y y los vectores aleatorios $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$, respectivamente. Para efectos prácticos, se supone que todas las X_j son v.a. discretas con soporte finito. El proceso de aprendizaje consiste en encontrar una red apropiada B dado el conjunto D.

En la Figura 2 se muestra un ejemplo de BN bajo el contexto de CS. El gráfico indica que, dado el nivel de ingresos del cliente, el perfil de riesgo (variable Clase) y el nivel de estudios son condicionalmente independientes. Similarmente, cuando el perfil de riesgo es conocido, el atributo cuenta bancaria es condicionalmente independiente del resto del resto de sus antecesores: nivel de estudios, ingresos, edad y sexo. Sin embargo, el riesgo de un cliente depende de manera causal directa de su nivel de ingresos, edad y sexo. Por su parte, los atributos nivel de estudios, edad y sexo son marginalmente independientes. Sin embargo, cuando los ingresos son conocidos, los dos primeros resultan ser condicionalmente dependientes, al

tiempo que cuando la clase es dada, los dos últimos son también condicionalmente dependientes.

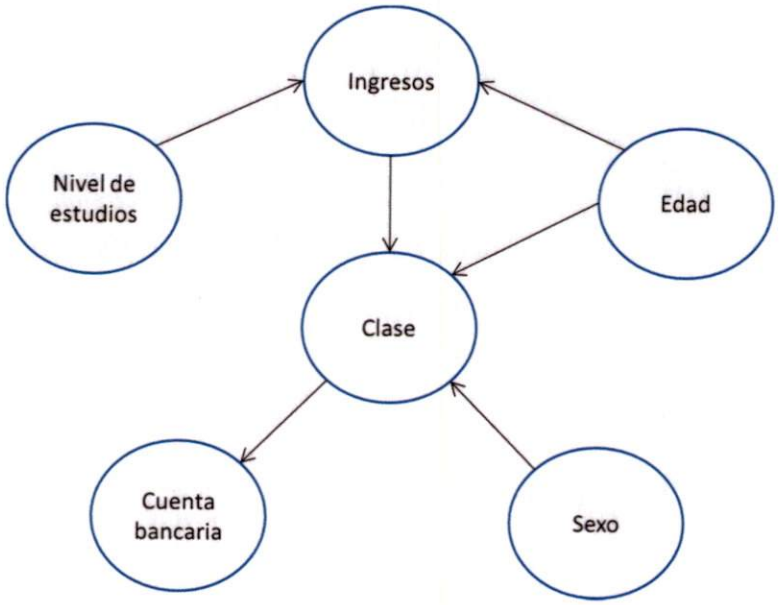

Figura 2. Ejemplo de una red bayesiana en el contexto de CS.

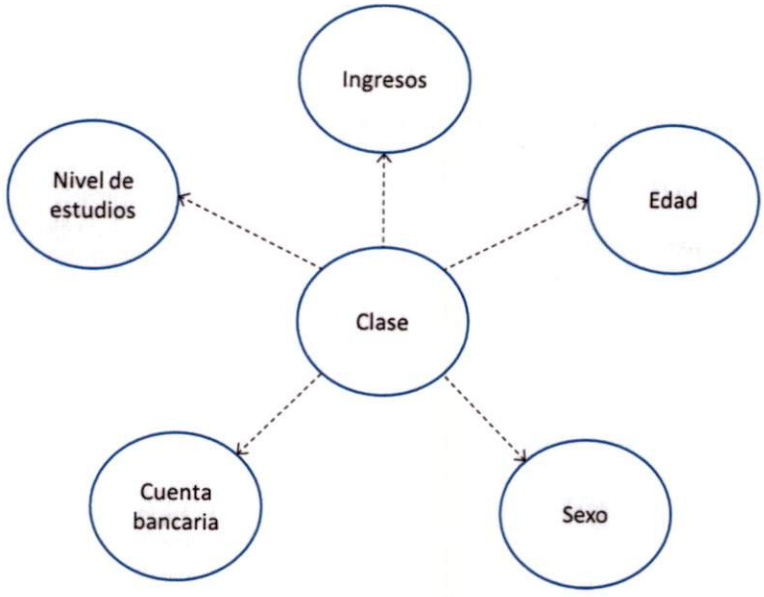

Figura 3. Ejemplo de la estructura de un clasificador *naive Bayes*.

Si consideramos los mismos atributos y variable de clase de la Figura 2, el clasificador *naive Bayes* para el problema de clasificación correspondiente, tendría la estructura presentada en la Figura 3. Como se puede observar, los arcos y nodos del gráfico satisfacen la ecuación (24).

Para utilizar una BN como clasificador, simplemente es necesario calcular $\text{argmax}_y P_B(y|\boldsymbol{x})$ empleando (24), la distribución P_B inducida por la red.

Ya que

$$P_B(Y|\boldsymbol{X}) = P_B(U)/P_B(\boldsymbol{X}) \propto P_B(U), \tag{26}$$

donde $U = \{Y, \boldsymbol{X}\}$, no es necesario complicar el proceso de inferencia que resulte de calcular $P_B(\boldsymbol{X})$.

La naturaleza dual de las BN divide a su proceso de aprendizaje en dos etapas. Primero se estima una estructura de red G, y enseguida se estiman las tablas de probabilidad, es decir, los parámetros Θ.

Algunas aproximaciones para realizar el aprendizaje de la estructura son:

1) Métricas de puntaje local. Se emplea alguna métrica de calidad $Q(B|D)$ la cual se desea maximizar. Entre las métricas utilizadas se encuentran aquellas construidas con aproximaciones bayesianas, verosimilitud, criterio de información de Akaike (AIC), longitud de descripción mínima (MDL), entropía y otros criterios de información. La propiedad de descomposición en la suma de puntajes individuales para cada nodo que presentan dichas métricas, permite el uso de métodos de búsqueda local.

2) Pruebas de independencia local. Este enfoque supone que existe una estructura de red que representa de manera exacta las independencias distribucionales de las v.a. subyacentes a los datos. Si se identifica independencia condicional entre dos nodos, esto indica que no es necesario conectarlos mediante algún arco.

3) Métricas de puntaje global. El desempeño de la BN sobre cierto conjunto de datos se mide estimando alguna métrica de utilidad esperada. Por ejemplo, se puede estimar la tasa de error de la red, o alguna otra medida de precisión, mediante validación cruzada[23]. A diferencia de las métricas de puntaje local, en este caso es difícil realizar una descomposición en puntajes individuales para cada nodo.

Entre los algoritmos de búsqueda comúnmente utilizados para implementar los enfoques mencionados arriba, se encuentran: *Hill climbing*, recocido simulado (*Simulated Annealing*), búsqueda tabú, algoritmos genéticos, simulación MCMC (*Markov Chain Monte Carlo*), TAN (*Tree Augmented Naive Bayes*) entre otros.

Una vez que se ha determinado una estructura para la BN, la cual por cierto, también pudiera suponerse prestablecida, es necesario llevar a cabo la estimación de las tablas de probabilidad condicional para cada nodo.

Un algoritmo simple para el aprendizaje de las BN es el denominado K2, el cual inicia con un orden prestablecido de los atributos. Se procesa cada nodo

[23] En el capítulo 4 se describen distintas alternativas para medir el desempeño de los modelos y métricas posibles, las cuales pueden ser empleadas también en el contexto de BN.

considerado y gradualmente considera la incorporación de arcos conectando nodos previamente procesados con el actual. En cada iteración se incorpora el arco que maximice el puntaje de la red.

En el caso en que la estructura de la BN tiene como único padre del resto de los nodos a la clase, el clasificador resultante es el naive Bayes, pues en este caso, todos los atributos son (condicionalmente) independientes dada la clase. Una extensión al clasificador simple de bayesiano propuesta por Friedman et al. (1997) es el denominado Tree Augmented Naive Bayes (TAN), los cuales relajan el supuesto de independencia condicional permitiendo la existencia de arcos entre los atributos. En las redes TAN, la variable de clase no tiene padres y cada atributo tiene como padre a la clase y a lo más alguno de los otros atributos. En consecuencia, los atributos forman una estructura de árbol.

El proceso de aprendizaje de las BN requiere un alto costo de conteo. Para cada estructura de red considerada en el algoritmo de búsqueda, los datos deben ser leídos en cada iteración para la construcción de las tablas de probabilidad condicional. Como alternativa, se ha propuesto almacenarlas en una estructura de datos en vez de eliminarlas en cada paso. Una estructura conocida es el árbol de todas dimensiones (*AD tree*).

3.5 Árboles de decisión

Muchos algoritmos con árboles de decisión han sido sugeridos previamente en la literatura. Uno de los más populares es el algoritmo C4.5 (Quinlan, 1993), el cual induce árboles de decisión basados en los conceptos de entropía y ganancia de información. Si se denota por p_c a la proporción de ejemplos de clase $c \in \{0,1\}$ en la muestra $S \subset D$. La entropía de S se define como

$$Entropía(S) = -p_1 log_2(p_1) - p_0 log_2(p_0)$$

donde $p_0 = 1 - p_1$. El valor máximo que la entropía de una muestra puede tomar es 1 (cuando $p_1 = p_0 = 0.5$) en tanto que su valor mínimo lo alcanza en 0 (cuando $p_1 = 0$ o $p_0 = 0$).

Se define la ganancia información de S con respecto al atributo x_i como la reducción en entropía por filtrar (dividir) los ejemplos sobre todos los valores de x_i

$$Ganancia(S, x_i) = Entropía(S) - \sum_{v \in valores(x_i)} \frac{|S_v|}{|S|} Entropía\ (S_v) \qquad (27)$$

donde S_v representa una submuestra de S en la cual el atributo x_i tiene un valor específico v. Cuando el criterio (27) es empleado para decidir hasta qué nodo dividir, el algoritmo favorece la división sobre atributos con muchos valores distintos. Por lo tanto, cuando un conjunto de datos contiene un atributo con un valor distinto para cada ejemplo (como por ejemplo, el caso de un campo identificador), el criterio de ganancia de información lo seleccionará como el mejor

criterio de división. Para solucionar esto, el algoritmo C4.5 aplica una normalización y emplea como métrica la razón de ganancia definida como

$$RazonGanancia(S, x_i) = \frac{Ganancia(S,x_i)}{Div(S,x_i)}, \tag{28}$$

con

$$Div(S, x_i) = -\sum_{k \in valores(x_i)} \frac{|S_k|}{|S|} log_2 \frac{|S_k|}{|S|}, \tag{29}$$

donde S_k representa una submuestra de S donde el atributo x_i tiene un valor específico k. La cantidad (29) representa la entropía de S con respecto de los valores de x_i. El uso del criterio (28) indica que C4.5 favorece divisiones sobre atributos que producen la mayor razón de ganancia bajo la restricción adicional de que la ganancia de información debe ser al menos tan grande como la ganancia promedio sobre todas las divisiones consideradas. El árbol entonces es construido por medio de partición recursiva. Esta estrategia de crecimiento en forma de árbol resulta entonces en un árbol complejo con muchos nodos intermedios que sobre ajustan a los datos. El algoritmo C4.5 plantea remediar esta situación por medio de un procedimiento de "poda", el cual se ejecuta retrospectivamente una vez que todo el árbol ha terminado de crecer. Un árbol de C4.5 "no podado" se puede traducir fácilmente en un conjunto de reglas mediante la construcción de reglas individuales para cada camino que vaya desde la raíz del árbol "no podado" hasta un nodo final ("hoja"). Estas reglas pueden ser posteriormente "podadas" al remover las condiciones basadas en un procedimiento similar al del árbol. El

algoritmo utiliza un parámetro de confianza para establecer el nivel de "poda". Valores pequeños de indican mayor nivel de "poda".

En la Figura 4 se muestra la estructura típica de un árbol de decisión para decidir si una nueva aplicación de crédito se aceptada o rechazada.

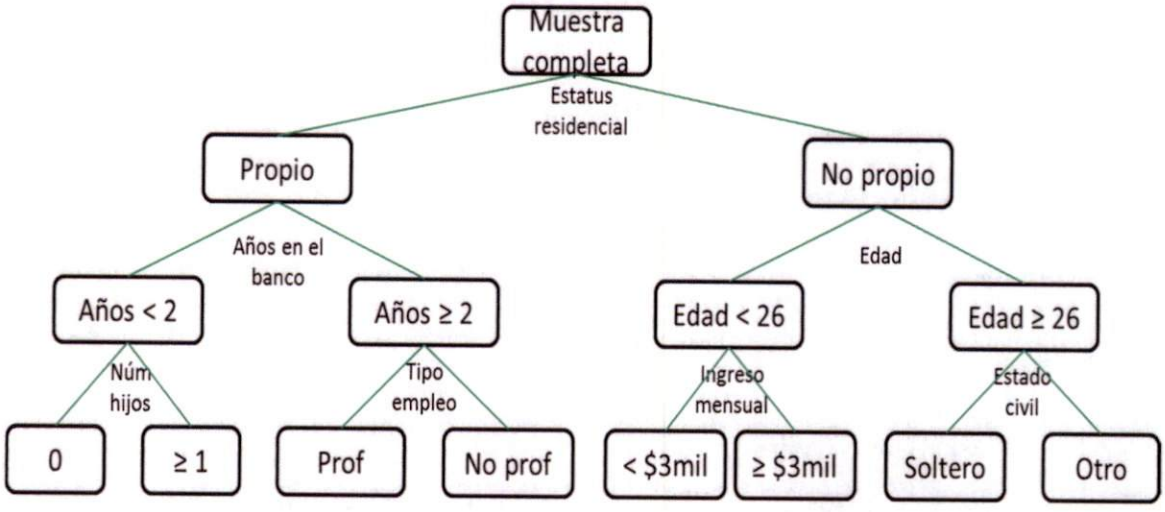

Figura 4. Árbol de decisión típico en CS.

Uno de los principales atractivos de los árboles de clasificación es su traducción directa en reglas de decisión que resultan fácilmente entendibles para usuarios que pudieran no estar muy familiarizados con el tema de técnicas de clasificación. Sin embargo, una de sus desventajas más importantes es su inestabilidad, ya que pequeños cambios en los datos pueden llevar a desviaciones importantes en las estimaciones, al tiempo que pueden sobre ajustar los datos si no se utiliza validación cruzada y poda.

3.6 Clasificador de k-vecinos más cercanos

El método de k-vecinos más cercanos (KNN) fue aplicado por primera vez en el área de CS por Chatterjee y Barcun (1970) y posteriormente por Henley y Hand (1996). La estrategia de este clasificador consiste en elegir una métrica sobre el espacio del conjunto de datos al que se desea aplicar para medir qué tan distante se encuentra un ejemplo de otro (Alpaydin, 2010). Con base en un conjunto de entrenamiento dado, un nuevo ejemplo es clasificado con buen o mal perfil de riesgo dependiendo de las proporciones de buenos y malos que mantengan los k-vecinos más cercanos al nuevo ejemplo. En consecuencia, esta técnica de clasificación depende básicamente de tres parámetros: la métrica, el número de instancias k que constituyen el conjunto de vecinos más cercanos, y la proporción de buenos que éstos debieran mantener para que el nuevo ejemplo sea clasificado como bueno.

La selección de la métrica es crucial. Fukanaga y Flick (1984) introdujeron una métrica general de la forma

$$d(\boldsymbol{x_i}, \boldsymbol{x_j}) = (\mathbf{x}_i - \mathbf{x}_j)\boldsymbol{A}(\mathbf{x}_i)\left((\mathbf{x}_i - \mathbf{x}_j)^{\mathbf{T}}\right)^{\frac{1}{2}} \qquad (30)$$

donde $\boldsymbol{A}(\mathbf{x})$ es una matriz de $p \times p$ simétrica definida positiva. $\boldsymbol{A}(\mathbf{x})$ se llama métrica local si depende de $\mathbf{x}$, y se denomina métrica global si es independiente de $\mathbf{x}$. Como la métrica local toma características específicas del conjunto de entrenamiento que pueden no ser apropiadas en general, muchos autores utilizan métricas globales. Henley y Hand (1996) utilizaron mezclas de distancias

euclidianas para separar clientes buenos de malos. Específicamente, estos autores sugieren una métrica de la forma

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left((\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\boldsymbol{I} + D\boldsymbol{w} \cdot \boldsymbol{w}^T)(\mathbf{x}_i - \mathbf{x}_j)\right)^{\frac{1}{2}} \tag{31}$$

donde $\boldsymbol{I}$ es la matriz identidad. La elección de los parámetros k y D se realiza mediante validación cruzada. En la literatura se han propuesto también medidas avanzadas de distancia para mejorar el desempeño de estos clasificadores.

El clasificador de vecinos más cercanos, aunque no es tan ampliamente utilizado en el área de CS como la regresión logística y el discriminante lineal, representa una alternativa útil. Probablemente una de las desventajas del método, desde el punto de vista práctico, es que esta técnica no permite construir un score para las características particulares de cada cliente.

3.7 Máquinas de Boltzmann restringidas

Una máquina de Boltzmann restringida (RBM) es una red neuronal con dos capas, una formada por unidades visibles y otra con unidades ocultas. En esta investigación, tanto las unidades ocultas como las visibles son consideradas como variables aleatorias binarias.

Una RBM es un tipo particular de las denominadas máquinas de Boltzmann (BM). Las BM fueron introducidas por Hinton et al. (1984) para referirse a un método no determinista diseñado para resolver un problema de visualización de imágenes. De

acuerdo con los autores, una BM es una red de unidades binarias estocásticas acopladas simétricamente. Contiene un conjunto de unidades visibles $\boldsymbol{v} \in \{0,1\}^D$, y un conjunto de unidades ocultas $\boldsymbol{h} \in \{0,1\}^D$. El número de capas ocultas y visibles están dados por n_h y n_v, respectivamente. La energía del estado $\{\boldsymbol{v}, \boldsymbol{h}\}$ se define como

$$E(\boldsymbol{v}, \boldsymbol{h}; \theta) = -\frac{1}{2}\boldsymbol{v}^t \boldsymbol{L}\boldsymbol{v} - \frac{1}{2}\boldsymbol{h}^t \boldsymbol{J}\boldsymbol{h} - \boldsymbol{v}^t \boldsymbol{W}\boldsymbol{h}, \tag{32}$$

donde $\theta = \{\boldsymbol{W}, \boldsymbol{L}, \boldsymbol{J}\}$ son los parámetros del modelo, los cuales representan las interacciones simétricas entre unidades visibles a ocultas, visibles a visibles y ocultas a ocultas, respectivamente. Los elementos de la diagonal de las matrices $\boldsymbol{L}$ y $\boldsymbol{J}$ son cero. El modelo asigna al vector visible $\boldsymbol{v}$ la probabilidad dada por

$$p(v; \theta) = \frac{1}{Z(\theta)} \sum_h e^{-E(\boldsymbol{v}, \boldsymbol{h}; \theta)} \tag{33}$$

donde el término normalizador $Z(\theta) = \sum_v \sum_h e^{-E(\boldsymbol{v}, \boldsymbol{h}; \theta)}$ se denomina función de partición. Las probabilidades condicionales para las unidades ocultas y visibles están dadas, respectivamente, por

$$p(h_j = 1 | \boldsymbol{v}, h_{-j}) = \sigma(\sum_i W_{ij} v_i + \sum_{k \neq j} J_{jk} h_j), \tag{34}$$

$$p(v_i = 1 | \boldsymbol{h}, v_{-i}) = \sigma(\sum_j W_{ij} h_j + \sum_{k \neq i} L_{jk} v_i), \tag{35}$$

donde $\sigma(x) = (1 + \mathrm{e}^{-x})^{-1}$ es la función logística. El algoritmo original para BM (Hinton y Sejnowski, 1983) suponía la inicialización aleatoria de cadenas de Markov para aproximar sus distribuciones de equilibrio y estimar una serie de operadores de esperanza matemática resultado de aplicar gradiente descendiente a la log-verosimilitud asociada al modelo de BM. El algoritmo desarrollado por estos

autores resultó muy lento para fines prácticos, por lo que ha sido necesario hacer simplificaciones como las planteadas en RBM.

A diferencia de BM, en RBM las conexiones solamente se permiten entre unidades de distintas capas pero no entre unidades de la misma capa. En la Figura 5 se muestran las arquitecturas típicas de BM y RBM.

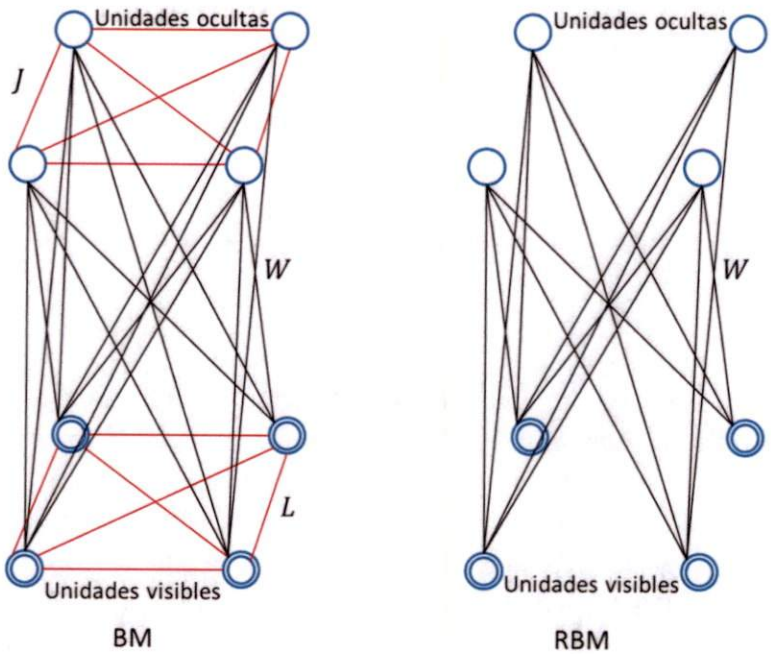

Figura 5. Arquitectura de BM (izquierda) y RBM (derecha)

La razón principal que motiva el uso de la arquitectura RBM es que se han desarrollado algoritmos de entrenamiento eficientes, como por ejemplo, el llamado algoritmo de divergencia contrastiva (CDA) propuesto por Hinton (2002).

A la especificación de RBM se le asocia una función de energía definida por

(36)

donde W es una matriz de pesos simétrica que conecta las unidades visibles (v) y ocultas (h), mientras que a y b son los vectores de sesgo para las capas visible y oculta, respectivamente.

Se puede verificar que toda RBM tiene la propiedad de que las variables visibles son condicionalmente independientes de las variables ocultas, y viceversa, específicamente

$$p(h_j = 1 \mid v) = \sigma\left(\sum_i w_{ij} v_i + a_j\right), \tag{37}$$

$$p(v_i = 1 \mid h) = \sigma\left(\sum_j w_{ij} h_j + b_i\right). \tag{38}$$

Los parámetros W, a, b se pueden estimar aplicando el algoritmo CDA (ver Hinton, 2002).

Para efectos de utilizar RBM como un clasificador, generalmente se emplea algún clasificador simple que entrene sobre la capa oculta. Para efectos de esta investigación, se utiliza la regresión logística para implementar la clasificación en el contexto de CS.

La información y búsquedas realizadas en la literatura hasta el momento, indican que no existe antecedente sobre el uso de RBM para crear clasificadores en el área de CS.

3.8 *Bagging* y *Boosting*

En el área de aprendizaje maquinal existen métodos de clasificación que combinan distintos clasificadores, de tal forma que el desempeño o capacidad predictiva del

conjunto completo (*ensemble*) es superior a la que posee cada uno de lo algoritmos de clasificación individualmente (Witten et al., 2011). En vez de elegir al clasificador con el mejor desempeño sobre el conjunto de entrenamiento, este enfoque sugiere utilizar los esquemas de clasificación disponibles, y posiblemente con desempeño individual bajo, para construir clasificadores híbridos. Una de las desventajas de esta aproximación es la interpretación de sus resultados, sin embargo, existen alternativas que permiten obtener descripciones estructuradas sobre los mecanismos del aprendizaje de estos métodos.

Entre los enfoques más prominentes para construir clasificadores combinando clasificadores se encuentran: *bagging*, *boosting* y *stacking*. La mayoría de las veces, estos esquemas han probado mejorar la capacidad predictiva con respecto a los clasificadores individuales. Los dos primeros enfoques utilizan un esquema de ponderación (votación) para construir las predicciones de clase. En el enfoque de *bagging*, los modelos reciben el mismo peso, mientras que en *boosting* los modelos con mejor desempeño son recompensados con ponderaciones mayores que el resto. Generalmente, estos enfoques se basan en el uso de árboles como unidades básicas.

En clasificación se consideran dos fuentes de error: 1) el sesgo (*bias*) o tasa de error, el cual mide qué tan bien el clasificador ajusta al problema concreto y sobre el cual considera un margen de error pues en la práctica es desconocido y solamente se puede aproximar; y 2) la varianza[24]. La suma de estas dos fuentes se denomina

[24] La fuente de error por varianza se debe al hecho de que el conjunto de entrenamiento es finito y en consecuencia puede no ser representativo de toda la población de ejemplos.

error total esperado del clasificador. Los métodos de *ensemble* buscan que con la inclusión de más clasificadores, se tenga una reducción en el componente de varianza.

Bagging busca neutralizar la inestabilidad de los métodos de aprendizaje alterando el conjunto de entrenamiento mediante la eliminación de algunas instancias y la réplica de otras, en vez de seleccionar una nueva muestra independiente en cada iteración. Las instancias se seleccionan aleatoriamente con reemplazo a partir del conjunto original para crear un nuevo conjunto del mismo tamaño, lo cual replica algunos ejemplos y elimina otros. El término bagging, que se utiliza para referirse a un proceso de *bootstrap* agregado, aplica algún esquema de aprendizaje específico sobre cada uno de los conjuntos artificialmente generados y los clasificadores generados a partir de éstos votan sobre la predicción de la clase. No obstante que los conjuntos construidos artificialmente no son independientes entre sí, generalmente mejora el desempeño en comparación con el esquema de aprendizaje individual y no presenta deterioros sustanciales. De hecho, *bagging* ayuda sustancialmente en el caso en que los esquemas de aprendizaje base sean inestables ante pequeños cambios en los datos de entrada. Por ejemplo, utilizar *bagging* sobre árboles de decisión sin proceso de poda ayuda a obtener mecanismos con mayor generalización debido a la inestabilidad de tales árboles. Una mejora a *bagging* se logra cuando los clasificadores utilizados generan probabilidades de clase en sus salidas, con lo que el esquema combinado también generaría probabilidades

ponderadas en su salida. *Bagging* también se ha utilizado bajo esquemas sensibles a costos.

Por su parte, *boosting* combina modelos múltiples capaces de complementarse entre sí, es decir, se explotan cada modelo en su parte del dominio donde mejor desempeño tiene respecto del resto. Al igual que *bagging*, este enfoque utiliza votación (para clasificación) o promedio (para predicciones numéricas) para combinar las salidas de los modelos individuales. Combina modelos del mismo tipo, como los árboles de decisión, por ejemplo. A diferencia de *bagging* donde los modelos individuales se construyen de forma separada, el esquema de *boosting* es iterativo, de manera que el desempeño de los modelos en cada iteración es influenciado por el desempeño de los modelos construidos previamente. Asimismo, pondera a cada modelo por su nivel de confianza en vez de utilizar pesos uniformes. Existen diversas variantes del enfoque *boosting*. Una de las más conocidas y específicamente diseñadas para clasificación es la denominada *AdaBoost.M1* (Witten et al., 2011). Este enfoque utiliza un peso específico para cada instancia, con lo que el error total de la clasificación se ve modificado como la suma de los pesos de todas las instancias mal clasificadas dividida por el peso de todas las instancias. El uso de pesos específicos sobre las instancias permite que el clasificador se concentre en ciertos subconjuntos de instancias. Inicialmente, el peso asignado a cada instancia es el mismo, sin embargo, conforme se realizan más iteraciones, las ponderaciones cambian para dar más peso a las instancias mal clasificadas, con lo que el algoritmo dedica más esfuerzo en clasificar las instancias

más "difíciles". Se monitorean los efectos causados sobre las instancias bien clasificadas en pasos anteriores y sobre las instancias cuya clasificación se ha podido corregir. El algoritmo busca un equilibro en las ponderaciones hasta alcanzar mejoras sustanciales mediante ciertos esquemas de actualización de los pesos (ver Witten et al., 2011).

El enfoque de *stacking* no es tan utilizado como los dos anteriores debido a la dificultad que generalmente involucra su análisis teórico y no tiene una estrategia estandarizada para su implementación. Este método permite combinar clasificadores de distinto tipo, por ejemplo, un árbol de decisión con un clasificador *Naive Bayes*. De igual forma que en *bagging*, este mecanismo utiliza votación.

Existe una basta literatura sobre los métodos de *ensemble* y su uso en en contexto de CS ha ido en incremento en la literatura reciente. El término *bagging* fue acuñado por Breiman (1996), mientras que el enfoque *stacking* fue propuesto por Wolpert (1992).

En este capítulo se presentaron las bases teóricas y metodológicas de los distintos clasificadores que serán utilizados para realizar el ejercicio empírico con distintos conjuntos de información real de créditos. En el siguiente capítulo se discuten los distintos esquemas de evaluación y métricas asociadas que se pueden utilizar para comparar el desempeño de los clasificadores utilizados.

Capítulo 4

Evaluación de los modelos de clasificación

En este capítulo se revisan algunas de las distintas métricas de desempeño de clasificadores, las cuales se han empleado, tanto en el área de minería de datos como de estadística, y particularmente en los problemas de clasificación y reconocimiento de patrones. Específicamente, se analizan diversas alternativas que tanto en la industria como en la academia, se han utilizado para comparar el desempeño de los modelos de calificación crediticia (CS).

Para evaluar la capacidad predictiva de un modelo que ha sido calibrado utilizando datos históricos, es necesario medir el desempeño que tendrá en el futuro sobre datos que no han sido utilizados en la creación del modelo. Asimismo, cuando se cuenta con varios modelos que buscan resolver algún problema de pronóstico común, resulta necesario contar con algún criterio de decisión para seleccionar al mejor modelo de acuerdo a su poder predictivo.

Usualmente, en el proceso de estimación y evaluación de los modelos de clasificación intervienen tres conjuntos importantes: entrenamiento, validación y prueba, los cuales pueden formar una partición del conjunto de datos[25]. El primero de ellos se utiliza para calibrar el modelo, en tanto que el segundo sirve para probar diferentes configuraciones de los parámetros o selección de variables. Por su parte, el conjunto de prueba está formado por instancias (también llamadas ejemplos u observaciones) no evaluadas durante el entrenamiento, el cual se utiliza para evaluar el modelo ajustado.

Como métrica natural del modelo ajustado, generalmente se calcula la tasa de error. La cual se puede entender como la probabilidad de que la clase actual y la predicción de clase obtenida bajo el modelo sean distintas. En la práctica, ésta se estima como el número de casos en que la clase fue mal pronosticada dividido sobre el número total de ejemplos o instancias.

Cuando se busca medir la calidad predictiva o de clasificación del modelo se recomienda hacerlo sobre el conjunto de prueba, ya que, por construcción, la tasa de error obtenida sobre el conjunto de entrenamiento tiende a ser menor que la tasa de error de un conjunto de prueba. Esto se debe principalmente a que el modelo ajustado es construido mediante algoritmos que minimizan el error de clasificación sobre los conjuntos de entrenamiento y/o validación. En consecuencia,

[25] El conjunto resultante de la unión de estos tres conjuntos puede coincidir con el conjunto de datos original, o bien, formar un subconjunto propio el cual puede ser obtenido mediante alguna técnica de muestreo algún otro criterio de segmentación.

la tasa de error no constituye una base de comparación adecuada. Asimismo, dos algoritmos no se pueden comparar con base al comportamiento de los errores obtenidos durante el entrenamiento, ya que es muy probable que el modelo con mayor número de parámetros o aquél con mayor complejidad, presente el mejor desempeño[26]. Estas observaciones nos indican que el proceso de validación del desempeño de un modelo debe realizarse con un conjunto distinto al utilizado en el entrenamiento.

Cuando las etapas de entrenamiento y prueba son realizadas en un solo conjunto, respectivamente, se corre el riesgo de obtener resultados que puedan estar afectados por ruido, observaciones atípicas (outliers) y otros factores aleatorios de las muestras. Para ello, es deseable repetir el experimento un cierto número de veces y promediar los resultados obtenidos. En el caso en que el costo de repetir los procesos de entrenamiento y prueba es demasiado alto, el ejercicio se puede realizar una sola ocasión.

La evaluación del algoritmo de aprendizaje se puede enfocar en el estudio del comportamiento de los errores de validación del modelo (sobre diferentes conjuntos de prueba), cuya distribución se puede comparar con la de los errores de validación de algún modelo alternativo. Sin embargo, de acuerdo con Turney (2000), existen

[26] Esto último sucede debido a que las construcciones de modelos más complejos generalmente se formulan sobre la base de modelos simples, ya sea heredando algunas propiedades de estos últimos, mezclando sus características o aumentando el número de parámetros para obtener generalizaciones de los mismos.

otros criterios que también se deben considerar para efectos de comparación entre modelos:

- Las implicaciones de generalizar la medida de error con funciones de pérdida[27].
- Tiempo de los procesos de entrenamiento y prueba, respecto de la complejidad del espacio paramétrico[28].
- Nivel de interpretación humana de las reglas de decisión inducidas por el modelo.
- Fácil implementación.

Finalmente, un aspecto que conviene tener presente, es que las conclusiones obtenidas del algoritmo de aprendizaje son condicionales al conjunto de datos. Como se ha comentado, el conjunto de entrenamiento se utiliza para estimar los parámetros bajo cierto algoritmo de aprendizaje, en tanto que el conjunto de validación generalmente se emplea para calibrar algunos parámetros específicos del algoritmo o de su estructura misma. Con el conjunto de prueba se determina el error del modelo que fue calibrado (entrenado) con la información contenida en los conjuntos de entrenamiento y validación.

---

[27] Una función de pérdida es una función que mapea al error del modelo en un número real que intuitivamente representa el "costo" asociado. Algunas funciones comunes son el valor absoluto y la función cuadrática.

[28] Se entiende por espacio paramétrico al subconjunto que contiene a todos los valores permisibles (posibles) que pueden tomar los parámetros del modelo.

4.1 Diseño y análisis de experimentos en los modelos de aprendizaje maquinal

En esta sección se discuten algunos conceptos, técnicas y estrategias relacionadas con el diseño y análisis de experimentos. Estos constituyen una fuente de principios estadísticos que han sido utilizados en el contexto del aprendizaje maquinal para comparar el desempeño de modelos y algoritmos.

Como primer paso, conviene establecer algunas definiciones relacionadas con los términos: experimento, factores, respuesta y estrategia de experimentación. Se define un experimento como una prueba o serie de pruebas las cuales tienen asociado a un conjunto de variables denominados factores. Los factores afectan directamente a los resultados de salida de un modelo de aprendizaje, el cual previamente ha sido entrenado sobre cierto conjunto de datos. Las variables sobre las cuales se puede tener control directo se denominan factores controlables, entre las que se encuentran, por ejemplo, el algoritmo utilizado, los parámetros, el conjunto de datos y sus atributos de entrada.

Existe otro tipo de factores llamados factores no controlables los cuales agregan variabilidad no deseada al proceso, pueden afectar las decisiones bajo el modelo, y no se tiene control directo sobre su comportamiento. Algunos ejemplos de este tipo de factores son: los componentes de ruido [29] en los datos, el subconjunto de

[29] El término ruido se refiere al efecto que tienen ciertos componentes aleatorios asociados con algunas de las variables o atributos del conjunto de datos. Por ejemplo, los modelos de regresión quedan definidos por una función que depende del conjunto de variables independientes o explicativas y el término de error aleatorio o ruido.

entrenamiento (en el caso de realizar muestreo de un conjunto más grande) y la aleatoriedad o los valores iniciales implícitos en el proceso de optimización.

Se denomina variable de respuesta a una cantidad construida a partir de los resultados de salida del modelo o algoritmo, con la que comúnmente se busca identificar a los factores más importantes. Como variable de respuesta se pueden utilizar: el promedio del valor absoluto del error de clasificación construido sobre el conjunto de prueba, o bien, alguna medida basada en alguna distribución de pérdida. Ejemplos específicos de medidas de desempeño para comparar distintos clasificadores se discuten en secciones posteriores.

Finalmente, la estrategia de experimentación se refiere al mecanismo o diseño del experimento requerido para derivar conclusiones del modelo estadísticamente significativas. En el caso particular del aprendizaje maquinal, se buscan modelos con la más alta capacidad de generalización (la menor tasa de error posible sobre el conjunto de prueba), mínima complejidad (bajo costo de implementación, en tiempo y espacio) y robustez (mínima afectación por fuentes de variabilidad externas). Una de las alternativas de experimentación es la denominada estrategia de mejor candidato, en la cual, dada una configuración inicial de candidatos, se analiza el comportamiento de la respuesta utilizando un candidato (o un subconjunto pequeño de estos) a la vez, probando las distintas combinaciones hasta que se llega a un buen estado. Este enfoque funciona bien cuando el experimentador cuenta con cierta intuición del proceso, sin embargo, no existe ningún criterio sistematizado para la selección de factores y tiempo de paro, es

decir, no hay garantía de encontrar la mejor configuración. Una estrategia alternativa es conocida como de un factor a la vez, la cual consiste en modificar algún factor específico manteniendo fijos los demás. Una desventaja notable de este enfoque es que no considera la interacción entre factores. Por otro lado, existe una estrategia llamada diseño factorial, en la cual los factores varían conjuntamente, razón por la cual resulta ser un enfoque de experimentación más robusto que los dos mencionados anteriormente, pero bajo un costo computacional mayor. Si se tienen F factores y L niveles, el enfoque de un factor a la vez toma un tiempo del orden $O(LF)$, mientras que el diseño factorial consume un tiempo $O(L^F)$.

Como primer paso para la implementación de experimentos de aprendizaje maquinal, es necesario definir objetivamente el propósito del estudio. Por ejemplo, se puede estar interesado en evaluar el error esperado o alguna otra medida de respuesta y evaluar si éste se encuentra en algún nivel aceptable. Por otro lado, es posible que se deseen comparar dos algoritmos de aprendizaje (o bien, el mismo algoritmo pero con alguna variante de mejora) sobre un conjunto de datos específico y se desea determinar aquél con mayor capacidad de generalización. Las comparaciones se pueden realizar sobre más de un conjunto de datos, o bien, tratarse de varios algoritmos que se desean comparar. En segundo lugar, se debe precisar la variable de respuesta a utilizar como medida de calidad de los modelos a comparar. Entre las alternativas de variable de respuesta más frecuentemente utilizadas se tiene la tasa de error de clasificación, el error cuadrático medio de una regresión, el uso de alguna distribución de pérdida para construir alguna métrica de

riesgo, o incluso incorporar alguna función de costo sobre las medidas utilizadas. Es necesario, por último: elegir los factores y niveles a utilizar (incluso si se considera alguna transformación o normalización de los niveles de los factores); el tipo de diseño experimental a implementar (se recomienda el diseño factorial a menos que se tenga certeza sobre interacción cero entre factores), las réplicas del experimento que se consideran así como tamaño de los subconjuntos; el desarrollo del experimento debe contemplar reproducibilidad de los resultados obtenidos, preferir el uso de código probado e incluso optar por técnicas de calidad en el desarrollo de software; finalmente, los experimentos deben garantizar el análisis estadístico de los datos y resultados obtenidos en los experimentos para derivar conclusiones significativas y con un sustento teórico.

Para concluir este apartado, se describen tres principios básicos sobre diseño de experimentos que se pueden aplicar sobre los conjuntos de entrenamiento y validación:

- Aleatorización. Consiste en un proceso de selección de muestras o réplicas de experimentos aleatoriamente determinados, de manera que los resultados obtenidos en cada iteración sean independientes.
- Réplicas. Tiene por objetivo lograr que para la misma configuración de factores controlables, el experimento se repita un cierto número de veces para luego promediar sobre el efecto de los factores no controlables. En el caso específico de aprendizaje maquinal, este efecto se logra corriendo el mismo algoritmo sobre un cierto número de subconjuntos del mismo conjunto de datos obtenidas

mediante algún tipo de muestreo. Esta técnica se denomina validación cruzada y permite evaluar la variabilidad de la respuesta por el efecto de los factores no controlables. Bajo este enfoque, es posible determinar umbrales del error experimental bajo los cuales se tienen conclusiones estadísticamente significativas.

- Bloques. Tiene por objetivo reducir o eliminar la variabilidad producida por factores cuya influencia sobre la variable de respuesta no nos interesa conocer. En el área de aprendizaje maquinal, cuando se realiza muestreo sobre cierto conjunto de datos, es necesario asegurar que las comparaciones de distintos modelos se realice sobre la misma base de subconjuntos muestrales. De no ser así, se puede incurrir en diferencias en desempeño de los modelos a causa a de experimentos realizados sobre distintos subconjuntos de prueba. En el contexto estadístico, los análisis realizados sobre la variable de respuesta bajo el enfoque de bloques se conoce como pruebas pareadas.

4.2 Validación cruzada y métodos de muestreo

Una manera estándar para evaluar los sistemas de CS consiste en utilizar una muestra tomada de la base de datos de clientes cuyo comportamiento fue observado en el pasado. Idealmente, esta muestra debiera ser independiente del subconjunto de datos utilizado para calibrar el modelo pero similar a la población sobre la cual se desea utilizar el sistema. En principio, esto genera un conflicto debido a las

tendencias en el comportamiento de la población ya que las características de los clientes pueden cambiar con el tiempo. Los nuevos clientes sobre los que se desea utilizar el sistema de calificaciones pueden ser completamente distintos a aquellos con los que el modelo se construyó y probó. Sin embargo, en muchas situaciones se dispone de información limitada para construir el modelo (se puede tratar de un nuevo producto o bien un nuevo grupo de clientes objetivo). En tal caso, se debe aprovechar mejor la información disponible (limitada), realizando subdivisiones de forma alternativa.

En este apartado se describe una de las alternativas de muestreo por réplicas más utilizado en aprendizaje maquinal: validación cruzada, así como otras alternativas de muestreo denominadas *leave-one-out* y *boostrapping*.

Una vez que parte del conjunto de datos se ha separado del mismo para ser utilizada como conjunto de prueba, se requiere construir parejas de conjuntos de entrenamiento y validación a partir del resto, el cual se denota por X y está formado por N instancias. En el caso en que X es un conjunto suficientemente grande, es posible dividirlo en una partición de tamaño K, de manera que cada una de estas se divida en dos, y una se utilice para entrenamiento y la otra para validación. Los valores de K que comúnmente se utilizan fluctúan entre 10 y 30, sin embargo, los conjuntos de datos pueden son ser los suficientemente grandes para llevar a cabo este tipo de partición.

4.2.1 Validación cruzada

Un método ampliamente utilizado en el cual se busca hacer el mayor aprovechamiento del conjunto de datos X mediante el uso de subconjuntos traslapados que conducen a errores dependientes. Se busca generar K parejas de conjuntos de validación/entrenamiento $\{T_i, V_i\}_{i=1}^{K}$ sobre el mismo conjunto X, de manera que el número de parejas distintas sea tan grande como sea posible, al tiempo que las estimaciones del error sean robustas y los traslapes entre conjuntos diferentes sean lo más pequeño posible. Por otro lado, se debe buscar que la proporción de ejemplos de cada clase permanezca invariante con respecto a la composición del conjunto de datos original (estratificación).

Más específicamente, la validación cruzada consiste en subdividir al conjunto X en K partes de igual tamaño, denotadas por $\{X_i\}_{i=1}^{K}$. Para construir el conjunto de validación se toma cualquiera de ellas y con el resto de $K-1$ partes se forma el conjunto de entrenamiento. Este proceso se repite para obtener K parejas de la forma $\{T_i, V_i\}$, donde $V_i = X_i$ y $T_i = \bigcup_{j \neq i} X_j$ para $i = 1,2,\ldots,K$.

La validación cruzada presenta dos problemas principales como consecuencia de su esquema de construcción. El primero tiene que ver con el alto grado de traslape entre conjuntos, ya que, por ejemplo, cualesquiera dos conjuntos de entrenamiento comparten $K-2$ particiones. Esto implica efectos de dependencia serial entre los errores obtenidos por cada pareja. El segundo problema tiene que ver con el hecho

de que para mantener un tamaño grande del conjunto de prueba, el tamaño del conjunto de validación debe ser pequeño.

A medida que el tamaño de K aumenta, el tamaño del conjunto de entrenamiento crece y entonces se obtienen estimadores del error más robustos, sin embargo, el conjunto de validación reduce cada vez más su tamaño. Adicionalmente, el proceso de entrenamiento incurre en un costo que están relación directa con el valor de K. A medida que N crece, el valor de K puede ser menor; mientras que si N es pequeño, los niveles que debe tomar K deben ser los suficientemente grandes para permitir contar con suficientes conjuntos de entrenamiento.

De acuerdo con Witten et al. (2011), se han realizado diversos estudios intensivos sobre numerosos conjuntos de datos, con distintas técnicas de aprendizaje, los cuales han mostrado que alrededor del nivel $K = 10$ se obtienen los mejores estimadores del error. Este caso se conoce como validación cruzada con 10 particiones (*10-fold*). Sin embargo, existe muy poca evidencia teórica que sustente este hecho. No obstante que estos argumentos no son concluyentes y el debate en las áreas de minería de datos y aprendizaje maquinal sobre cuál es el mejor esquema para la evaluación sigue vigente. La validación cruzada con 10 particiones se ha posicionado como el método estándar en términos prácticos. Los autores también afirman que los ejercicios empíricos encontrados muestran que el uso de estratificación mejora ligeramente los resultados de estimación del error, debido a la ganancia que se obtiene por la reducción en varianza (aunque ésta no se elimina completamente). Adicionalmente, una estrategia para mejorar la estimación del

error consiste en repetir 10 veces el proceso de validación cruzada *10-fold*, para finalmente promediar los resultados obtenidos. Esto evidentemente crea una demanda de recursos computacionales importante[30].

4.2.2 *Leave-one-out*

Este es el caso extremo de validación cruzada en donde se toma $K = N$, es decir, dado un conjunto con N instancias, se elige un elemento para formar el conjunto de validación y el resto de las $N - 1$ instancias se utilizan como conjunto de entrenamiento. El elemento que se ha dejado fuera del conjunto de entrenamiento se evalúa con una función indicadora indicando si fue o no correctamente clasificado bajo el modelo ajustado con el conjunto de entrenamiento. El promedio de los N resultados representa la estimación final del error.

Este procedimiento presenta dos principales atractivos. Primero, el proceso de entrenamiento utiliza la mayor cantidad de datos, con lo que presumiblemente se espera aumentar el nivel de clasificación del modelo y proporciona estimaciones del error lo más robustas posible. En segundo lugar, como esta técnica es determinista (a diferencia de validación cruzada en donde el proceso de muestreo puede arrojar distintos resultados) se obtiene siempre el mismo resultado en la estimación del error.

[30] Otras elecciones que también han mostrado resultados positivos son valores de K entre 5 y 20.

En contraparte, este procedimiento presenta un alto costo computacional ya que el proceso de entrenamiento del clasificador debe ser ejecutado N veces, lo cual usualmente es poco factible cuando se tiene un conjunto de datos muy grande. Adicionalmente, por construcción, este esquema no permite estratificación en los conjuntos obtenidos (es imposible mantener las proporciones de clases del conjunto original cuando se tiene sólo una observación en el conjunto de validación).

4.2.3 *Bootstrapping*

Es un método de estimación del error del modelo basado en un procedimiento estadístico denominado muestro con remplazo. En este procedimiento, una instancia puede aparecer más de una vez en los conjuntos de entrenamiento. Del conjunto X se seleccionan N elementos, con reemplazo, para formar otro conjunto R con N instancias que se utiliza como conjunto de entrenamiento. Por su parte, el conjunto completo se utiliza como conjunto de validación. Las muestras obtenidas por este procedimiento se traslapan aún más que en el caso de validación cruzada y en consecuencia los errores obtenidos presentan mayor dependencia; sin embargo, es considerada la mejor manera de hacer muestreo con conjuntos pequeños. En el contexto específico de CS, el modelo ajustado sobre el conjunto obtenido por el muestreo con reemplazo del conjunto original se suele evaluar sobre este último.

La probabilidad de que una instancia específica no sea seleccionada después de N intentos en el muestreo con reemplazo está dada por $\left(1-\frac{1}{N}\right)^N \approx e^{-1} \approx 0.368$, para

N suficientemente grande. Esto significa que el conjunto de entrenamiento incluye aproximadamente al 63.2% de las instancias de X. Como el sistema no será entrenado en el 36.8% de los datos, la estimación del error sobre el conjunto de validación será muy pesimista. Una solución inmediata consiste en repetir el proceso, digamos m veces para generar conjuntos $R_1, \ldots, R_m$ con los que se analiza el comportamiento de los errores obtenidos. Para ofrecer una corrección en su estimación del error, el procedimiento de *bootstrapping* combina la tasa de error $error(X)$ obtenida en el conjunto de validación con los errores de sustitución $error(X - R_i)$ asociados al conjunto de entrenamiento. Entonces la tasa de error se calcula como

$$TE = 0.368 * error(X) + 0.632 * \sum_{i=1}^{m} error(X - R_i) \qquad (39)$$

El conjunto $X - R_i$ representa a las instancias que están en X pero no en R_i.

4.3 Métricas para el desempeño del clasificador

En este apartado se abordan distintas métricas que sirven para comparar el desempeño de varios clasificadores.

En el caso de los problemas de clasificación, la información básica sobre el desempeño del modelo ajustado se puede resumir en la denominada matriz de confusión. Para la situación de un problema de dos clases, los componentes de dicha matriz se pueden describir de la manera siguiente:

		Clase verdadera	
		Positivo	Negativo
Clase pronosticada	Positivo	Número de positivos verdaderos (TP)	Número de positivos falsos (FP)
	Negativo	Número de negativos falsos (FN)	Número de negativos verdaderos (TN)

Figura 6. Matriz de confusión

De acuerdo con la Figura 6, las entradas sobre la diagonal representan el número de aciertos que tiene el modelo cuando se clasifica un cierto conjunto de ejemplos u observaciones de cierta población de los cuales se conoce su verdadera clase. El objetivo de la matriz de confusión consiste en contabilizar cuatro cantidades importantes: el número de ejemplos correctamente clasificados como clase "positivo" (TP, por sus siglas en inglés: *True Positive*) y los correctamente clasificados en la clase negativo (TN, por sus siglas en inglés: *True Negative*), así como el número de ejemplos de la clase "positivo" que fueron clasificados erróneamente como "negativo" (FN, por sus siglas en inglés: *False Negative*) y los casos clasificados como "negativo" pero que pertenecen a la clase "positivo" (FP, por sus siglas en inglés: *False Positive*). La generalización de esta matriz para el caso de problemas de más de dos clases también es útil para contabilizar el total de aciertos y errores cometidos por el modelo en el pronóstico de la clase a la que pertenece cada ejemplo probado (Witten et al., 2011).

Para los problemas de clasificación donde el número de clases es $d > 2$ se pueden diseñar matrices de confusión similares a la Figura 5. En este caso, la matriz de confusión se construye como una matriz de dimensión $d \times d$ cuyas entradas (i, j) contiene el número de instancias que pertenecen a la clase C_i pero han sido asignadas a la clase C_j. En el caso ideal en que no se tiene error de clasificación, todas las entradas fuera de la diagonal debieran ser igual a cero. Alternativamente, se pueden definir d problemas de dos clases, cada uno separando una clase específica de las $d - 1$ clases restantes. Para el contexto de este trabajo, en el cual se estudia el problema de CS, inicialmente el análisis se enfoca en clases binarias (buenos y malos clientes). Sin embargo, en situaciones en las que se requiera definir una escala de calificaciones múltiples, que contemple distintas categorías de clientes, se recomienda el uso de matrices de confusión de múltiples clases.

Con la información contenida en la matriz de confusión es común construir diversas métricas de desempeño, entre las que destacan:

$$Tasa\ de\ éxito = \frac{TP+TN}{TP+TN+FP+FN} \tag{40}$$

$$Tasa\ de\ error = 1 - Tasa\ de\ éxito \tag{41}$$

$$Tasa\ de\ positivos\ verdaderos = \frac{TP}{TP+FN} = Recall = Sensibilidad \tag{42}$$

$$Tasa\ de\ positivos\ falsos = \frac{FP}{FP+TN} = Fallout \tag{43}$$

$$Precisión = \frac{TP}{TP+FP} \tag{44}$$

$$Especificidad = \frac{TN}{TN+FP} = Tasa\ de\ negativos\ verdaderos \quad (45)$$

$$Medida\ F = \frac{2}{\frac{1}{Precisión} + \frac{1}{Sensibilidad}} \quad (46)$$

$$Media\ harmónica = \frac{2*Sensibilidad*Especificidad}{Sensibilidad+Especificidad} \quad (47)$$

En las expresiones anteriores se han incluido los nombres alternativos que en distintos contextos se han utilizado para referirse a la misma métrica. En la práctica y la literatura se pueden encontrar métricas adicionales, o bien, variantes de las mismas. Para expresar las medidas anteriores como tasas en %, es necesario multiplicar cada cantidad por 100.

La tasa de éxito, también conocida como porcentaje de clasificación correcta (TC) de las observaciones, mide la proporción de los casos correctamente clasificados en una muestra de datos. En un cierto número de casos, el TC puede que no sea el mejor criterio de desempeño. Tácitamente asume que los costos de mala clasificación son equivalentes para predicciones positivamente falsas y negativamente falsas. Este supuesto es problemático, ya que para la mayoría de los problemas en la vida real, un tipo de error de clasificación puede resultar más caro que otro. Otro supuesto implícito del uso del TC como una forma de criterio de evaluación es que la distribución de clases (a priori de la clase) entre los ejemplares, se presume constante en un horizonte de tiempo y es relativamente balanceado. Además, usar el TC por sí solo, a veces demuestra ser inadecuado toda vez que las distribuciones de clases y los costos de mal calificar clientes son raramente

uniformes. De cualquier forma, tomar en cuenta las distribuciones de clases y los costos de mal calificar clientes parece ser un poco duro, ya que en la práctica, éstos raramente pueden ser especificados con precisión y en general se encuentran sujetos a cambios.

La sensibilidad mide la proporción de ejemplos positivos que se predijo iban a ser positivos (es decir, $TP/(TP + FN)$), mientras que la especificidad mide la proporción de ejemplares negativos que se predijo iban a ser negativos (o sea, $TN/(FP+TN)$). Note que sensibilidad, especificidad y el TC varían juntos conforme el umbral en una salida continua de un clasificador varía entre sus extremos.

4.3.1 Receiver operating characteristics curve (ROC)

En el contexto del problema de dos clases, consideremos un sistema que devuelve la probabilidad de que una instancia pertenezca a la clase positiva, la cual denotamos por $P(C_1|x)$. Entonces la probabilidad asociada a la clase negativa está dada por $P(C_2|x) = 1 - P(C_1|x)$. Adicionalmente, supongamos además que $P(C_1|x) > \theta$. Entonces, para θ cercano a 1, resulta poco probable que dada una instancia arbitraria, ésta pertenezca a la clase positiva; es decir, se no habría positivos falsos sino unos pocos positivos verdaderos. A medida que θ disminuye su valor acercándose a cero, el número de positivos verdaderos aumenta pero se incurre en el riesgo de introducir positivos falsos en el análisis de clasificación.

Para distintos valores de se pueden obtener las parejas ordenadas de positivos verdaderos y positivos falsos. Al conjunto de puntos que se obtiene la llamada curva o gráfica ROC[31] (*Receiver Operating Characteristics*) como se muestra en la Figura 7. La curva ROC es una representación gráfica bidimensional de la *sensibilidad* (en el eje Y) vs la cantidad *1-especificidad* (en el eje X) para distintos valores de la clasificación del umbral. Esto básicamente ilustra el comportamiento de un clasificador sin tener en cuenta la distribución de las clases o el costo de mal calificar clientes, de modo que desacopla efectivamente el desempeño de la clasificación de estos factores.

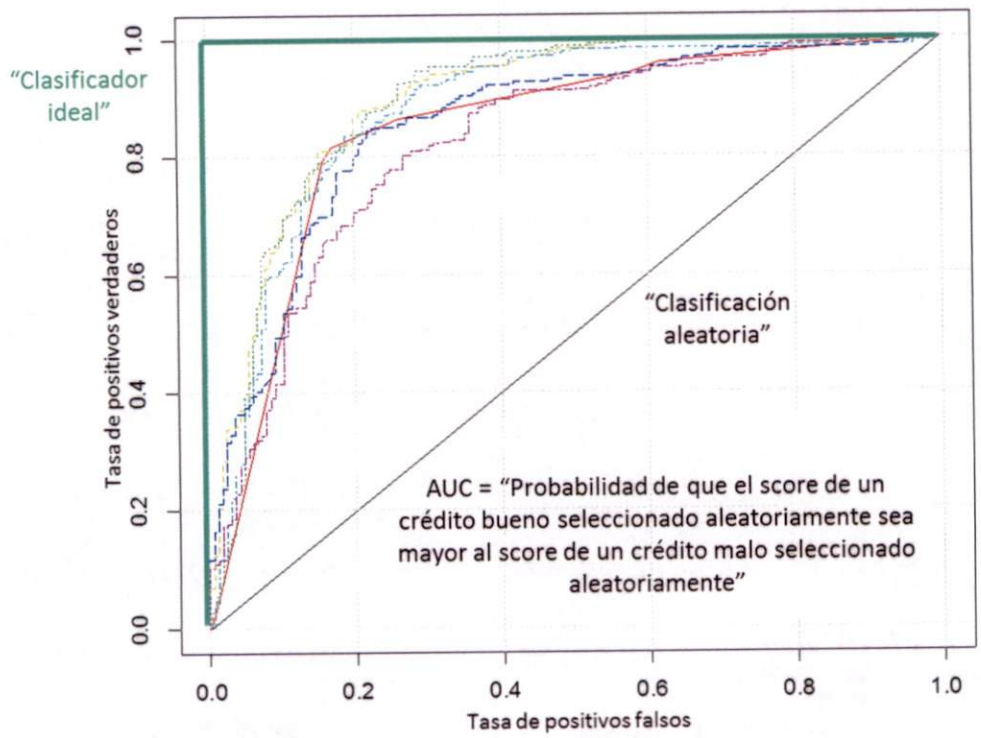

Figura 7. Curva ROC

[31] También conocida como diagrama de Lorenz (Thomas et al., 2002).

En el caso ideal de un clasificador con cero tasa de error, se tendría una tasa de positivos verdaderos igual a 1, mientras que la tasa de positivos falsos sería 0. Es decir, la forma de ROC sería la representada por la línea verde de la Figura 7. La situación ilustrada por la recta identidad (pendiente 1) corresponde a un clasificador que arrojar el mismo número de decisiones verdaderas que falsas, lo cual representaría el peor escenario de desempeño de cualquier clasificador (en el caso de estar por debajo de esta recta bastaría con invertir el orden de las clases para regresar a un comportamiento como el de la curva en rojo). Dados dos clasificadores, se puede decir que uno de ellos es mejor que otro si la curva ROC del primero se encuentra por encima de la del segundo. En el caso en que ambas curvas se intersecten, el criterio de decisión común suele ser el área debajo de ROC (AUC, por sus siglas en inglés). El clasificador ideal tendría AUC igual a 1 y distintos clasificadores se comparan de acuerdo a su AUC comparando así su desempeño promedio sobre distintas condiciones de valores de funciones de pérdida[32]. Alternativamente, la máxima distancia entre la curva ROC y la recta identidad define a la estadística KS (*Kolmogorov-Smirnov*) asociada a este problema. El valor de KS también ha sido utilizado como criterio de comparación entre modelos. En el contexto de CS, al área AUC está directamente relacionado

[32] Se puede demostrar que distintos valores de θ corresponden a diferentes matrices de pérdida para ambos tipos de error. En consecuencia, la curva ROC se puede considerar como el comportamiento del clasificador bajo distintas matrices de pérdida (Alpaydin, 2010).

con el denominado coeficiente de Gini[33]. Un aspecto importante en este contexto tiene que ver con la utilidad que tiene ROC para determinar puntos de corte para discriminar a los buenos de los malos clientes (Thomas et al., 2002).

Para efectos de este trabajo, se utilizarán a TC (la tasa de clasificación correcta) y AUC (el área bajo ROC) como métricas de desempeño de los modelos calibrados.

Para el caso de mútiples clases, Hand y Till (2001) propusieron una generalización de la métrica AUC. Por otro lado, Hand (2009) propuso como alternativa al uso de AUC una medida coherente denominada medida H. En la literatura existen otras métricas alternativas como, por ejemplo, las llamadas gráficas de costos propuestas por Drummond y Holte (2006).

El contenido de este capítulo se centró en la discusión de distintos esquemas de evaluación del desempeño de clasificadores en el contexto del CS. De entre las alternativas consideradas, en el siguiente capítulo se empleará la validación cruzada para estimar las métricas de desempeño seleccionadas. Se considera un ejercicio de experimentación en el cual se repite cierto número de veces el entrenamiento de los clasificadores bajo validación cruzada, para posteriormente aplicar alguna prueba estadística que permite establecer si existen diferencias significativas en el nivel de desempeño de los distintos modelos empleados.

---

[33] El coeficiente de Gini se define como el doble del área formada entre la curva ROC y la recta identidad. Un clasificador ideal tendría un valor de Gini igual a 1, mientras que un clasificador aleatorio tendría un valor asociado de 0. Tanto la estadística KS como el coeficiente de Gini y AUC, miden el desempeño del clasificador sobre todo el rango de posibles puntos de corte. Sin embargo, bajo el contexto de CS, generalmente el interés se centra en solamente un rango pequeño de puntos de corte.

Capítulo 5

Resultados de los clasificadores sobre los conjuntos de datos

5.1 Análisis descriptivo

En esta sección se realiza un análisis descriptivo de los conjuntos de datos los cuales se utilizarán posteriormente para calibrar modelos de scoring sobre la base de ciertos métodos estadísticos tradicionales y de aprendizaje maquinal.

Se cuenta con acceso a tres bases de datos públicas y dos de carácter privado, cuyo origen no se proporciona por razones de confidencialidad. La información pública contiene dos conjuntos de datos disponibles en el repositorio de UCI (http://archive.ics.uci.edu/ml/). Uno de ellos, denominado "German credit data", corresponde a información sobre distintos atributos de 1000 personas clasificadas con un buen o mal perfil de riesgo de crédito. El segundo conjunto "Australian credit approval" contiene 690 instancias de aplicaciones de crédito las cuales están

clasificadas como aceptadas o rechazadas. Por su parte, el tercer conjunto público corresponde a información proveniente de la operación de tarjetas de crédito de una gran cadena de retail en Brasil durante el periodo 2003-2008. Dicha información fue solicitada en la página de la 13th Pacific-Asia Knowledge Discovery and Data Mining conference (http://sede.neurotech.com.br:443/PAKDD2009/), en la cual se proporcionan dos subconjuntos, uno de entrenamiento formado por 50,000 instancias y otro para prueba constituido por 10,000 ejemplos. A este conjunto se le refiere como "PAKDD2009".

Con respecto a la información privada, se cuenta con un conjunto denominado "Cerveza" el cual contiene información relativa a 3,749 clientes a quienes una empresa vendió cerveza durante 2009 y 2010. Para 2,593 de estos clientes, la venta del producto se realizó mediante el otorgamiento de líneas de crédito las cuales fueron monitoreadas para incluir luego indicadores del comportamiento de su perfil de riesgo de los clientes a quienes se otorgó crédito. El segundo conjunto disponible, corresponde a información socio-demográfica y de seguimiento de patrones de pago de una cartera de 48,550 préstamos de auto otorgados entre 2006 y 2009 por una empresa mexicana especializada en el financiamiento automotriz [34] . Las características propias de este conjunto denominado "Autos", permite implementar

[34] La institución a la que se hace referencia, y cuyo nombre se omite por razones de confidencialidad, opera bajo la figura de sociedad financiera de objeto múltiple no regulada (SOFOM ENR). Este tipo de instituciones están sujetarse a las correspondientes disposiciones de la Ley General de Organizaciones y Actividades Auxiliares de Crédito, así como a las que emitan en los términos de dicha Ley la Comisión Nacional Bancaria y de Valores (CNBV) y la Secretaría de Hacienda y Crédito Público (SHCP). No obstante, las Sofomes ENR, no están sujetas a la supervisión de la CNBV.

un análisis estático del perfil de riesgo de los clientes (buenos/malos), al tiempo que permite sentar las bases para futuros estudios sobre el comportamiento dinámico del perfil de riesgo-rentabilidad de los clientes mediante técnicas de aprendizaje maquinal.

En la siguiente lista se anota entre paréntesis el nombre que se utilizará en el resto del documento para hacer referencia a cada conjunto:

- "German credit data" (Alemán)
- "Australian credit approval" (Australiano)
- "PAKDD2009" (PAKDD)
- "Cerveza" (Cerveza)
- "Autos" (Autos)

Los conjuntos Alemán y Australiano han sido ampliamente utilizados en la literatura para comparar distintas técnicas de clasificación, tanto en el área del credit-scoring como en contextos más generales. Por su parte, a raíz del reto PAKDD 2009, el conjunto PAKDD ha recibido atención en algunos artículos recientes. En cambio, la información de los conjuntos Cerveza y Autos representan información proveniente del mercado mexicano. El uso de conjuntos reales y que representen características de la economía local, permitirá evaluar los alcances de técnicas alternativas provenientes del aprendizaje maquinal para el caso mexicano.

A manera de resumen sobre las características de los distintos conjuntos se incluye la Tabla 1 (ver Apéndice). En el siguiente apartado se proporciona una descripción

de las características principales de cada conjunto de datos, así como el tipo de problema específico que cada uno permitirá abordar en el contexto de credit scoring y en línea con los objetivos planteados en este proyecto de investigación.

5.1.1 Conjunto Alemán

Este conjunto de datos contiene información relativa a 1,000 solicitantes de crédito, descritos por un conjunto de atributos. El archivo original proporcionado por el Dr. Hans Hofmann se encuentra compuesto por atributos cuyos valores están expresados como variables categóricas, en su mayoría, de acuerdo con la Tabla 2 (ver Apéndice).

El atributo de clase indica el perfil crediticio de cada solicitud mediante dos clases: bueno y malo. Sin incluir al atributo de clase, se tienen 20 atributos (7 numéricos y 13 nominales). Del total de ejemplos, 700 tienen buen perfil crediticio y 300 están clasificados con mal perfil.

Para trabajar con una versión del conjunto alemán con atributos numéricos, la universidad de Strathclyde, Glasgow, produjo un segundo archivo editado el cual cuenta con la inclusión de algunas variables indicadoras para su uso en algoritmos que no trabajan con variables categóricas. Algunos de los atributos de tipo categórico fueron codificados como enteros. El conjunto resultante cuenta con un total de 24 atributos (sin incluir la variable de clase) y, como se ha indicado, todos ellos son de tipo numérico.

La información disponible no indica la presencia de observaciones faltantes y no proporciona detalle sobre el procesamiento de los datos.

Los algoritmos utilizados para este conjunto permitirán clasificar los ejemplos en alguna de las dos clases existentes en función de los atributos disponibles. Este problema es de tipo estático y con él se busca determinar si con el análisis de los atributos se puede determinar el perfil de riesgo de crédito de nuevas solicitudes.

5.1.2 Conjunto Australiano

Contiene información sobre otorgamiento de tarjetas de crédito. Los nombres de los atributos y sus valores originales fueron transformados a ciertos símbolos con la finalidad de proteger la confidencialidad de los datos. El archivo original fue proporcionado por Quinlan (1986).

Este conjunto de datos contiene 690 ejemplos, de los cuales 307 pertenecen a la clase con mal perfil crediticio (malo) y 383 tienen un buen perfil de riesgo (bueno). Se tienen 14 atributos (sin incluir el atributo de clase), 6 de los cuales son numéricos y 8 categóricos. La Tabla 3 (ver Apéndice) incluye la descripción de cada uno de los atributos.

De acuerdo con la información proporcionada en el sitio de UCI, los datos faltantes se reemplazaron por el promedio observado en cada atributo en el caso de las variables de intervalo (numéricas) y por la moda del atributo en el caso de las variables categóricas.

Ambos conjuntos de datos, el alemán y australiano, serán utilizados para aplicar los modelos estadísticos tradicionales y algunas técnicas de aprendizaje maquinal para discriminar a los clientes de buen perfil (Buenos) de los clientes (Malos) en un enfoque estático.

Es importante mencionar que las dos fuentes de información han sido ampliamente utilizadas en la literatura para comparar distintos modelos de clasificación. Por mencionar algunos estudios recientes se tienen los trabajos realizados por Baesens et al. (2003), Hoffmann et al. (2007), Huang et al. (2006, 2007), Ong et al. (2005), West (2000, 2005), Zhou (2008).

5.1.3 Conjunto PAKDD

La información disponible corresponde a la utilizada en un desafío sobre minería de datos organizada por NeuroTech Ltd y el Centro de Informática de la Universidad Federal de Pernambuco, Brazil, presentada en 13th Pacific-Asia Knowledge and Dada Mining conference (PAKDD 2009).

Para propósitos de la competencia, se proporcionaron tres conjuntos de información[35]: entrenamiento, tabla de posiciones y prueba. Los tres conjuntos

[35] La forma en que opera este producto de crédito es estándar. Los clientes (aplicaciones de crédito aceptadas) utilizan su tarjeta para realizar compras en la cadena departamental, las cuales se facturan a la cuenta del cliente para pagar entre 10 y 40 días después de realizada la compra. La fecha límite de pago se fija en cierto día del mes. Un cliente se etiqueta como malo si, para una ventana de observación de 11 meses después de su primera compra, éste cuenta con algún incumplimiento de pago (morosidad mayor a 60 días). En otro caso el cliente es considerado como buen cliente. En el caso de algún consumo reciente, se incluyen 60 días más de observación como periodo de maduración de la última facturación.

contienen información capturada a lo largo de un año completo, cada uno durante periodos no adyacentes. El conjunto de entrenamiento consta de 50,000 ejemplos con los que se debe extraer información que mejor explique el perfil de riesgo de los clientes (bueno/malo), el cual incluye 9,874 clientes con mal perfil de riesgo y 40,126 con buen perfil.

Los otros dos conjuntos sirven para evaluar el desempeño de los modelos ajustados y cada uno cuenta con 10,000 ejemplos. Por razones de la dinámica de la competencia PAKDD 2009, para ninguno de estos dos conjuntos se provee la columna de clase. La intención del conjunto de tabla de posiciones es poder realizar un ejercicio previo, mientras que el conjunto de prueba se utilizó para evaluar las distintas metodologías implementadas con la finalidad de elegir a la ganadora.

Los ejemplos de cada conjunto fueron obtenidos mediante muestreo aleatorio de experiencia histórica registrada en portafolios de tarjetas de crédito otorgadas por una de las cadenas departamentales (retail) más grandes de Brasil. En cada caso, la base inicial contiene información sobre 31 variables explicativas, y solamente en el conjunto de entrenamiento, se dispone del atributo de clase (perfil de riesgo).

La compañía ha operado el otorgamiento de las tarjetas de crédito durante más de 8 años aplicando básicamente dos distintos modelos de evaluación del riesgo, de manera que su tasa de aceptación ha variado entre 50% y 75%. Asimismo, no obstante que las condiciones económicas del periodo en cuestión han mostrado cierta estabilidad, imperfecciones como el ruido, datos faltantes, informaciones

atípicas (outliers) y otros cambios en el mercado, pueden reducir el desempeño de los modelos estimados.

En la Tabla 4 del Apéndice se incluye la descripción de los atributos, en la que no se incluye al identificador (único) del cliente.

Para el conjunto de prueba se detectaron 8 atributos cuyos valores tienen tasa de faltantes del 100% (ya sea porque todos sus valores están vacíos o bien tienen un solo valor constante para todos los ejemplos), los cuales serán eliminados de la muestra para propósitos del análisis. Como resultado se mantienen 22 atributos (10 de tipo numérico y 22 no numéricas).

Se propone utilizar alguna transformación logarítmica para disminuir el efecto que tienen las observaciones con valores atípicamente grandes (outliers), así como considerar el uso de algún umbral máximo permisible.

5.1.4 Conjunto Cerveza

Este conjunto de datos consta de registros de clientes a los que se les vendió cerveza durante 2009 y 2010. La base de datos original consta de 3,749 clientes, de los cuales a 2,593 se les otorgaron líneas de crédito en función de ciertos atributos valorados por la empresa cervecera en cuestión[36], una de las cervecerías más importantes de México. Se disponen 18 atributos (2 numéricos y 16 nominales), de los cuales dos son de tipo clase: uno caracteriza el otorgamiento de línea de crédito

[36] Por razones de confidencialidad se ha omitido el origen de la fuente de datos

(Si/No) y el otro se describe el comportamiento del perfil de riesgo (bueno/malo). La autorización de líneas de crédito responde a una estrategia de la empresa de financiar a sus clientes para que aumente su capacidad de compra de cerveza y con ello lograr un mejor posicionamiento de la marca en el mercado.

En principio, el estudio busca identificar los patrones de aquellos clientes a quienes se otorgó línea de crédito. Asimismo, interesa discriminar de entre estos clientes aquellos que resultaron con buen comportamiento crediticio de los malos pagadores. La caracterización de los rasgos que mejor describen a cada tipo de cliente permitirá a la empresa realizar una mejor selección de clientes a los que es conveniente vender financiar mediante la asignación de líneas de crédito.

La Tabla 5 (ver Apéndice) contiene información sobre el tipo de variables contenidas en la base de datos original.

En un análisis exploratorio de los valores faltantes, se detectó que todos los clientes con valores faltantes en los atributos Esquina, Calle, Zona y Local correspondían a los mismos registros (505) y que además todos formaban parte de los faltantes en el atributo Imagen. Luego de eliminar tales registros, el número de casos con información nula en el atributo Clave se redujo de 291 a solamente 22, en tanto que para los atributos Imagen y Control se encontraron 3 y 10 casos, respectivamente. Finalmente, se eliminaron 4 clientes en los que el atributo Resultado reportó ventas totales negativas. Como resultado de omitir los filtros y consideraciones realizadas, el conjunto de datos a analizar contiene 3,200 clientes. Cabe destacar que de estos, solamente a 2,156 les otorgaron crédito.

En la Tabla 6 del Apéndice se muestra la distribución porcentual de los atributos categóricos (no numéricos), de acuerdo con los valores asignados a cada categoría. Por su parte, la Tabla 7 del Apéndice contiene estadísticas descriptivas para las únicas dos variables numéricas (de intervalo) del conjunto.

Como ya se ha comentado, este conjunto de datos permitirá realizar dos estudios principales. El primero consistirá en desarrollar modelos de clasificación que expliquen el proceso de otorgamiento de la línea de crédito a los 3,200 clientes totales en función de los primeros 14 atributos mostrados en la primera Tabla 6. El segundo, consistirá en evaluar el impacto de cada uno de tales atributos sobre el comportamiento crediticio de los 2,156 clientes a quienes se otorgó línea de crédito, es decir, sobre su buen perfil (cumplido) o mal perfil (incumplido).

5.1.5 Conjunto Autos

Este conjunto contiene información de 48,550 créditos de auto otorgados entre 2006 y 2009 por cierta institución financiera local.

En la Tabla 8 (ver Apéndice) se encuentra la descripción de los atributos sociodemográficos que se dispone de cada uno de los créditos otorgados en todo el periodo de la muestra. La base sociodemográfica cuenta con 24 atributos, de los cuales 17 son de tipo numérico y 7 nominales.

Para algunos de los atributos que presentan ejemplos con valores atípicamente grandes (outliers) se contempla el uso de transformaciones logarítmicas y la

eliminación de los casos a partir de cierto rango. Entre estos atributos se encuentran: número de dependientes, ingresos brutos, si cuenta con algún crédito hipotecario, ingresos netos, otros ingresos, deudas, precio del automóvil a financiar. Se cuenta además, con una base que contiene la información del comportamiento de pago observado en los clientes de la base sociodemográfica entre 2008 y 2010 (ver Tabla 9 del Apéndice).

Con la información contenida en la base de comportamiento de pago, será posible realizar distintos análisis del perfil de los clientes (buenos y malos) en función de las características sociodemográficas, la fecha de otorgamiento del crédito y comportamiento de pago de los clientes por meses de saldo vencido. Este tipo de análisis está alineado con el enfoque dinámico planteado en el proyecto de investigación.

5.2 Resultados de la experimentación

Con excepción del clasificador basado en RBM, las distintas técnicas empleadas fueron calibradas con ayuda del software Weka, el cual está disponible en http://www.cs.waikato.ac.nz/~ml/weka/index_downloading.html. Este software se ubica entre las herramientas más utilizadas para minería de datos[37]. Para el caso de RBM, los análisis se implementaron con apoyo de una librería en MATLAB, la cual se puede descargar de forma gratuita en http://code.google.com/p/matrbm/.

[37] El autor puede consultar, por ejemplo, la siguiente liga con los principales *software* de minería de datos que los usuarios han utilizado durante los últimos 12 meses para implementar proyectos reales: http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html

Finalmente, parte del análisis exploratorio y descriptivo de la información se realizó con el paquete estadístico R y la librería de minería de datos Rattle, ambos disponibles desde el sitio http://cran.r-project.org/.

El uso del software Weka facilitó la selección de valores adecuados para los parámetros principales a utilizar en cada clasificador (excepto RBM[38]) calibrado en cada uno de los conjuntos de datos descritos en la sección anterior. Para ello, se empleó validación cruzada para determinar las configuraciones de los valores de los parámetros de manera tal que en cada clasificador y conjunto empleado, se obtuviera la menor tasa de error de clasificación.

La Tabla 10 contiene los resultados del proceso de selección de parámetros bajo validación cruzada[39]. En dicha tabla, se ha incluido una columna que indica algunos aspectos importantes sobre las ventajas y desventajas de cada clasificador desde el punto de vista de la facilidad de interpretación de las salidas que cada modelo arroja para fines de clasificación. El término "Caja Negra" se utiliza para hacer referencia a situaciones en las que el funcionamiento del clasificador no es

[38] En el caso de RBM se realizó una búsqueda exhaustiva del número de capas ocultas en el rango de 1 a 500 mediante la librería en MATLAB citada arriba. El criterio de decisión empleado fue también la tasa de error bajo validación cruzada.

[39] El conjunto Cerveza Línea consiste de la muestra completa de clientes a quienes del conjunto Cerveza. Por su parte, el conjunto Cerveza 90+ corresponde al subconjunto de clientes de Cerveza Línea a quienes se otorgó línea de crédito para la venta de cerveza. En este último conjunto, al igual que en Autos 90+, se utilizó como criterio para definir un mal perfil crediticio (clientes incumplidos) a todos aquellos casos en donde la morosidad resultó superior a 90 días de mora (3 meses).

transparente para el usuario en el sentido de que es difícil capturar la contribución que cada atributo tiene para explicar la variable de clase.

Clasificador	Australiano	Alemán	Cerveza Línea	Cerveza 90+	PAKDD	Autos 90+	Ventajas/Desventajas
Regresión Logística	Ridge $(R) = 2.4242$	$R = 4.7959$	$R = 2.0408$	$R = 4.47$	$R = 2.2414$	$R = 3.6578$	Traducción de un score en probabilidades [Bajo]
Naive Bayes	Discretización supervisada (DS)	DS	DS	DS	DS	DS	Supuesto simple de independencia condicional [Bajo]
K-vecinos más cercanos	#vecinos más cercanos $(k) = 19$	$k = 17$	$k = 8$	$k = 13$	$k = 25$	$k = 21$	Simple/Difícil traducción a score de nuevos clientes [Bajo]
Árbol J48	Factor de confianza $(c) = 0.1189$	$c = .25$	$c = .4268$	$c = .01$	$c = .15$	$c = .7393$	Interpretación y entendimiento de reglas simples/ Sensible a pequeños cambios portafolio [Medio]
Red perceptrón multicapa	Capas/unidades ocultas $(n_h) = 8$; tasa de aprendizaje $(L) = .3$	$n_h = 6$; $L = .01$	$n_h = 3$; $L = .12$	$n_h = 5$; $L = .05$	$n_h = 6$; $L = .15$	$n_h = 10$; $L = .25$	Posibilidad de traducción en probabilidades score/ "Caja negra" [Alto]
Máquina soporte vectorial (SVM) base radial	Costo $(C) = 32768$; gamma $(\gamma) = .000488$	$C = 4096$; $\gamma = .00781$	$C = 2$; $\gamma = .25$	$C = 3.5$; $\gamma = .01$	$C = 256$; $\gamma = .125$	$C = 173$; $\gamma = .0274$	Posibilidad de traducción en probabilidades score; extracción de reglas/ "Caja negra" [Alto]
Red bayesiana	Discretización (D); Algoritmo de búsqueda local (BL) K2; No base Naive Bayes	D; BL=TAN	D; BL=TAN; Criterio Entropía	D; BL=Tabú	D; BL=Tabú	D; BL=TAN; Criterio Entropía	Uso de tabla de distribución de probabilidades condicionales [Regular]
AdaBoostM1	Clasificador base (CB) = Logística	CB = Logística	CB = Logística	CB = Logística	CB = Logística	CB = Logística	"Caja negra" [Alto]
Bagging	Clasificador base (CB) = Árbol de decisión rápido (REPTree)	CB = REPTree	CB = REPTree	CB = REPTree	CB = REPTree	CB = REPTree	"Caja negra" [Alto]
RBM + Logística	Unidades ocultas $(n_h) = 180$	$n_h = 90$	$n_h = 250$	$n_h = 170$	$n_h = 260$	$n_h = 110$	Traducción de un score a probabilidades/ "Caja negra" [Alto]

Tabla 10. Configuración de parámetros utilizados en los clasificadores para cada conjunto de datos

Finalmente, se incluye entre corchetes, y en escala categórica (alto, medio, bajo), al nivel de dificultad que cada método presentó durante el proceso de selección de parámetros[40].

Empleando la configuración seleccionada para los valores en los parámetros de los clasificadores, se realizó un análisis sobre los atributos que muestran mayor estabilidad en la explicación de la variable de clase. Nos obstante que no es todos los clasificadores es posible determinar el nivel de explicación o contribución que mantienen en el modelo, se construyó la Tabla 11 (ver Apéndice). Se incluyeron las variables que mantuvieron mayor nivel de significancia (en el caso de la regresión lineal), que aparecieron como atributos iniciales en las reglas de decisión inducidas por el clasificador de árboles (J48), o bien, que también sirvieron como nodos principales en las estructuras de redes bayesianas construidas. Excepto para el caso del conjunto Australiano (donde no se conoce la descripción de atributos) los atributos mostrados indican consistencia con las variables a las cuales típicamente se les da importancia en el proceso de calificación crediticia en la industria.

Posteriormente, se llevó a cabo un ejercicio de experimentación en el cual se repite 100 veces el ajuste de los clasificadores bajo validación cruzada (*10-fold*), realizando un total de 1,000 corridas o ajustes bajo los parámetros descritos en la Tabla 10.

---

[40] La determinación del nivel de dificultad como una variable categórica obedece al hecho de que la dificultad y esfuerzo para la calibración de sus parámetros en que cada método incurrió es una mezcla del tiempo computacional incurrido, el número de parámetros calibrados, los posibles rangos de valores del espacio de búsqueda y las mejoras relativas que cada refinamiento presentó respecto a la disminución en tasa de error.

Derivado de los análisis de los resultados, en la Tabla 12 se reportan los valores promedio de las medidas las tasas de clasificación correcta (TC) y el área bajo la curva ROC (AUC) de cada modelo[41].

Clasificador	Australiano		Alemán		Cerveza Línea		Cerveza 90+		PAKDD		Autos 90+	
	% TC	AUC	% TC	AUC	% TC	AUC	% TC	AUC	% TC	AUC	% TC	AUC
Red Bayesiana	86.35	0.92	74.56	0.77	82.83	0.87	60.85	0.65	81.14	**0.85**	**72.82**	**0.76**
NaiveBayes	85.28	0.92	74.55	0.77	81.93	0.84	60.21	0.65	80.75	0.84	70.56	0.72
LibSVM	85.54	0.86	75.65	0.77	84.15	0.87	60.81	0.65	81.44	0.73	71.48	0.72
Regresion Logistica	86.38	0.93	76.85	0.78	83.82	0.87	61.27	0.66	81.26	0.85	72.14	0.74
Perceptrón multicapa	83.38	0.90	75.47	0.78	83.10	0.87	58.94	0.63	78.97	0.80	70.87	0.69
Vecinos mas cercanos	86.17	0.92	72.49	0.75	83.65	0.89	61.87	0.66	**81.45**	0.82	69.23	0.70
AdaBoostM1	86.23	0.90	**76.92**	0.72	83.85	0.84	61.29	0.60	81.25	0.82	72.09	0.76
Bagging	86.32	**0.94**	76.29	**0.79**	**85.93**	**0.89**	60.55	0.65	81.32	0.82	71.23	0.75
Árboles (J48)	85.13	0.88	73.57	0.68	84.46	0.85	61.53	0.63	81.36	0.75	70.04	0.71
RBM+Logistica	**86.55**	0.87	75.07	0.69	85.02	0.88	**61.95**	**0.66**	80.23	0.83	72.59	0.75

Tabla 12. Tasas de clasificación correcta (TC) y área bajo ROC (AUC) de los clasificadores para cada conjunto de datos

Para evaluar si las diferencias obtenidas en el desempeño son significativas se utilizó una prueba T para muestras pareadas con factor de corrección

$$t = \frac{\bar{d}}{s_d\sqrt{\left(\frac{1}{n}+\frac{n_2}{n_1}\right)}} \tag{47}$$

El nivel de significancia utilizado es del 5% y se ha señalado en color rojo los casos en que se tiene el peor desempeño. En color azul se indican los casos en que los modelos resultaron tener un desempeño inferior al alguno de sus competidores. Asimismo, se ha puesto en color negro (sin formato) las situaciones en la cuales los modelos son igualmente competitivos entre sí (i.e., no se encontraron diferencias

[41] La naturaleza de los valores reportados, puede llevar a que las pruebas t sean no transitivas por diferencias significativas entre las varianzas muestrales de los resultados de experimentación en cada modelo y conjunto de datos.

significativas). Finalmente, los casos en negritas indican los clasificadores que presentaron los niveles más altos bajo la métrica de desempeño específica.

Se encuentra consistencia en los resultados obtenidos para los conjuntos de datos Alemán, Australiano y PAKDD

Los resultados indican que el clasificador híbrido *Bagging* mostró el mejor desempeño en tres conjuntos de datos (4 métricas con valores máximos), seguido por los clasificadores red Bayesiana y RBM+Logística con mejor desempeño en un par de conjuntos cada uno (3 métricas con valores máximos). Adaboost y Vecinos más cercanos presentaron el mejor desempeño para un conjunto de datos respectivamente. El perceptrón multicapa utilizado mostró los peores desempeños en cuatro de los conjuntos de datos, al tiempo que LibSVM y Vecinos más cercanos mostraron los peores desempeños, cada uno, en un par de conjuntos. Por su parte, el clasificador de árboles de decisión tuvo el peor desempeño para un conjunto, pero en el resto de los datos presentó siempre un desempeño inferior que alguno de los otros competidores. Finalmente, es conveniente resaltar que la regresión logística, aunque no mostró valores máximos en las métricas utilizadas, mantuvo el desempeño más estable sobre todos los conjuntos de datos al resultar un clasificador estadísticamente competitivo con respecto al resto de modelos.

Capítulo 6

Conclusiones y recomendaciones para trabajo futuro

Como parte de esta investigación, se ha realizado un análisis detallado sobre el desempeño de distintos clasificadores en el contexto del *credit-scoring* (CS). Para ello, se implementó un ejercicio de experimentación sobre distintos conjuntos de datos disponibles aplicando diversas técnicas de clasificación, algunas de las cuales pertenecen al área de estadística y otras han sido desarrolladas en el área del reconocimiento de patrones. Son cinco los conjuntos de información utilizados, tres de los cuales (Alemán, Australiano y PAKDD) se encuentran disponibles de forma pública mientras que los dos restantes (Cerveza y Autos) son de tipo privado. La información pública, representativa de otros mercados en el extranjero, ha sido utilizada en numerosos estudios de clasificación y CS, en tanto que la base privada pertenece al mercado local mexicano.

Por las características propias de cada conjunto de datos, con los distintos modelos implementados se analizó el perfil de riesgo, tanto de contrapartes que solicitan crédito, como de aquellos clientes previamente aceptados. Asimismo, los análisis empíricos realizados fueron orientados de forma intensiva para resolver el problema de CS bajo un contexto de tipo estático.

Entre las técnicas de clasificación empleadas se encuentran: regresión lineal, árboles de decisión, vecinos más cercanos, *naive Bayes*, redes perceptrón multicapa, máquinas de soporte vectorial, *boosting*, *bagging* y un clasificador híbrido formado por máquinas de Boltzmann restringidas (RBM) y regresión logística. Hasta donde se tiene conocimiento, este trabajo es pionero en el uso de esta última técnica para atacar el problema de clasificación en el área de CS. Los esfuerzos de este proyecto se enfocaron en explorar la aplicabilidad de tales metodologías sobre datos reales del mercado mexicano, evaluando las bondades y desventajas en cada caso. Específicamente, se sientan bases útiles para el uso de técnicas alternativas de clasificación en portafolios de créditos con las características propias del mercado mexicano, contribuyendo para futuros desarrollos en la materia.

Entre las principales conclusiones obtenidas de los modelos calibrados sobre los distintos conjuntos de datos, se tiene que no existe un clasificador cuyo desempeño sea superior sobre el resto de las metodologías implementadas en este documento[42]. Sin embargo, los resultados obtenidos con base en las medidas de desempeño

[42] Estas conclusiones son consistentes con un resultado desarrollado por Wolpert (1996) en el contexto de aprendizaje maquinal (*no free lunch theorem).*

utilizados (TC y AUC) indican que varios de los algoritmos pertenecientes al área de reconocimiento de patrones son tan competitivos como las técnicas estadísticas tradicionales. Más aún, los modelos híbridos *Bagging* y RBM + Logística mantuvieron los niveles de desempeño más alto, por lo menos en alguna de las dos métricas con las que se desarrolló el estudio.

Como futuro trabajo de investigación, se plantean dos líneas principales. La primera se refiere al estudio y aplicación de técnicas de clasificación aún no exploradas en el área del CS, como por ejemplo, las denominadas redes de aprendizaje profundo (DBN), las cuales están formadas a su vez por varias capas de RBM, entre otras alternativas relativas al área de *deep learning.* Como segunda línea de investigación, se plantea la posibilidad de explorar la aplicabilidad de algunos de los clasificadores usados en este documento, para realizar un análisis dinámico del comportamiento de los clientes en el tiempo. Sobre este último aspecto, se pueden precisar varios enfoques del estudio. En primer lugar, se puede citar el análisis de la evolución del perfil de riesgo de los clientes[43]. Para ello, técnicas como el análisis de supervivencia y modelos de Markov ocultos (HMM, por sus siglas en inglés) son un par de alternativas a considerar. Como segundo aspecto que resulta de gran importancia para la gestión de los portafolios de crédito, se considera la inclusión de componentes como la rentabilidad para extender la

[43] Este aspecto resulta en principio relevante para las instituciones que desean evaluar los niveles de pérdida futuros en el portafolio, producto del deterioro de la calidad crediticia de las contrapartes y los eventos de incumplimiento. Asimismo, es de gran interés para las compañías, cuantificar los impactos que sobre los requerimientos de capital tiene la evolución del riesgo de crédito en sus portfolios.

clasificación a distintos perfiles de clientes en del sentido rendimiento-riesgo que representan para las instituciones. Finalmente, un aspecto complementario a los dos puntos ya mencionados, se refiere a la construcción de escenarios del desempeño de los clientes con base en estrategias de mitigación de pérdidas (morosidad, cobranza y recuperación de cartera) que resulten de la detección de patrones y alertas tempranas bajo múltiples técnicas de clasificación.

Referencias

[1] Alpaydin, E. (2010). Introduction to Machine Learning. The MIT Press. 2da. Ed.

[2] Apilado, V.P., D.C. Warner y J.J. Dauten (1974). Evaluative techniques in consumer finance. Journal of Financial Quantitative Analysis, 9(2), 275-283.

[3] Alfo, M., S. Caiazza y G. Trovato (2005). Extending a Logistic Approach to Risk Modeling through Semiparametric Mixing. Journal of Financial Services Research, 28(1), 163-176.

[4] Altman, E.I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance, 23(4), 589–609.

[5] Altman, E.I. (2005). An emerging market credit scoring system for corporate bonds. Emerging Markets Review, 6, 311-325.

[6] Anderson, R. (2007). The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation. Oxford University Press.

[7] Antonakis, A.C. y M.E. Sfakianakis (2009). Assessing naive Bayes as a method for screening credit applicants. Journal of Applied Statistics, 36(5), 537-545.

[8] Arns, M.T., P.J. Steiner, N.Y. Soma, T. Shimizu y J.C. Nievola (2006). Using neural network rule extraction for credit-risk evaluation. International Journal of Computer Science and Network Security, 6(5A), 6-17.

[9] Atiya, A.F. (2001). Bankruptcy prediction for credit risk using neural networks: a survey and new results. IEEE Transactions on Neural Networks, 12(4), 929-935.

[10] Barniv, R., J.B. McDonald (1999). Review of Categorial Models for Classification issues in accounting and finance. Review of Quantitative Finance and Accounting, 13, 39-62.

[11] Bastos, J. A. (2008) Credit scoring with boosted decision trees. City: Munich Personal RePEc Archive, pp. 262-273.

[12] Baesens, B., M. Egmont-Petersen, R. Castelo y J. Vanthienen (2002). Learning Bayesian network classifiers for credit scoring using Markov Chain Monte Carlo search. In Proceedings of the 16th International Conference on Pattern, 3, 49-52.

[13] Baesens, B., R. Setiono, C. Mues y J. Vanthienen (2003a). Using neural network rule extraction and decision tables for credit risk evaluation. Computer Journal of Management Science, 49(3), 312-329.

[14] Baesens, B. T. Gestel, S. Viane, M. Stepanova, J. Suykens y J. Vanthienen (2003b). Benchmarking state of the art classification algorithms for credit scoring. Computer Journal of Operational Research Society, 54(3), 627-635.

[15] Banasik J., J.N. Crook y L.C. Thomas (1999). Not if but when borrowers default. Journal of Operational Research Society, 50, 1185-1190.

[16] Bicer, I., D. Sevis, y T. Bilgic (2010). Bayesian credit scoring model with integration of expert knowledge and customer data. En 24th Mini EURO Conference, 324-329.

[17] Borra, S. y S. Caiazza (2002). Comparative performance of credit scoring models using aggregated predictors. Data Mining III, WIT Press, 747-756.

[18] Breiman, L., Friedman, J.H., Olshen, R.A. y C.J. Stone (1984). Classification and Regression Trees. Wadsworth, Belmont, CA.

[19] Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140.

[20] Caiazza, S. (2004). The comparative performance of credit scoring models: an empirical approach. En Monetary Integration, Markets and Regulation, Research in Banking and Finance, 4, 17-66.

[21] Cessie, S. y J.C. Houwelingen (1992). Ridge estimators in logistic regression. Journal of Applied Statistics, 41(1), 191-201.

[22] Chatterjee, S. y S. Barcun (1970). A nonparametric approach to credit screening, Journal of the American Statistical Association, 65, 150-154.

[23] Chen, M.C. y S.H. Huang (2003). Credit Scoring and Rejected Instances Reassigning through Evolutionary Computational Techniques. Expert Systems with Applications, 24(4), 433–441.

[24] Coffman, J.Y. (1986). The proper role of tree analysis in forecasting the risk behavior of borrowers, MDS Reports, Management Decision Systems, Atlanta, GA, 3-9.

[25] Craven, M.W. y J.W. Shavlik (1996). Extracting tree-structured representations of trained networks. En Advances in Neural Information Processing Systems. MIT Press, Cambridge, MA, 24-30.

[26] Crook, J.N., D.B. Edelman y L.C. Thomas (2007). Recent developments in consumer credit risk assessment; European Journal of Operational Research. 183(3), 1447-1465.

[27] Cybenko, G. (1989). Approximations by superpositions of sigmoidal functions. Mathematics of Control, Signals, and Systems, 2(4), 303-314.

[28] Davis, R.H., D.B. Edelman, A.J. Gammerman (1992). Machine-learning algorithms for credit-card applications, IMA Journal of Mathematics applied in Business and Industry, 4, 43-52.

[29] Durand, D. (1941). Risk Elements. En Consumer instalment financing. National Bureau of Economic Research, New York. Disponible en http://www.nber.org/books/dura41-1

[30] Drummond, C. y R. Holte (2006). Cost curves: an improved method for visualizing classiffer performance. Machine Learning, 65, 95–130.

[31] Fan, A. y M. Palaniswami (2000). A new approach to corporate loan default prediction from financial statements. En Proc. Computational Finance/Forecasting Financial Markets Conf. CF/FFM-2000, London (CD), UK.

[32] Fantazzini, D. y S. Figini (2009). Random survival forests models for SME credit risk measurement. Methodology and Computing in Applied Probability 11(1), 29-45.

[33] Fisher, R.A. (1936). The Use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179–188.

[34] Fix, E. y J. Hodges (1952). Discriminatory analysis, nonparametric discrimination, consistency properties. Report 4, Project 21-49-004, School of Aviation Medicine, Randolph Field, TX.

[35] Friedman, N., D. Geiger y M. Goldszmidt (1997). Bayesian network classifiers. Machine Learning 29, 131–163.

[36] Fukanaga, K. y T.E. Flick (1984). An optimal global nearest neighbour metric. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-l, 25-37.

[37] Galindo, J. y P. Tamayo (2000). Credit risk assesment using statistical and machine learning: basic methodology and risk modelling applications. Computational Economics, 15, 107-143.

[38] Grablowsky, B.J. y W.K. Talley (1981). Probit and discriminant functions for classifying credit applicants: A comparison: Journal of Economics and Business, 33, 254-261.

[39] Ghodselahi, A. (2011). A hybrid support vector machine ensemble model for credit scoring. International Journal of Computer Applications, 17(5).

[40] Hand, D.J. y W.E. Henley (1997a). Statistical classification methods in customer credit scoring: a review. Journal of the Royal Statistical Society 160(3): 523–541.

[41] Hand, D. J., K. McConway y E. Stanghellini (1997b). Graphical models of applicants for credit. IMA Journal of Mathematics Applied in Business and Industry, 8(2), pp. 143–155.

[42] Hand, D.J. y M.G. Kelly (2001). Lookahead scorecards for new fixed term credit products. Journal of the Operational Research Society, 52, 989-996.

[43] Hand, D.J. y R.J. Till (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning, 45(2),171–186.

[44] Hand, D.J (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning, 77,103-123.

[45] Henley, W.E. y D.J. Hand (1996). A k-NN classifier for assessing consumer credit risk. Statistician, 65, 77-95.

[46] Hinton, G.E. y T.J. Sejnowski (1983). Optimal perceptual inference. En IEEE conference on Computer Vision and Pattern Recognition.

[47] Hinton, G.E., T.J. Sejnowski y D.H. Ackley (1984). Boltzmann machines: Constraint satisfaction networks that learn. Reporte técnico TR-CMU-CS-84-119, Carnegie-Mellon University, Dept. of Computer Science.

[48] Hinton, G.E. (2002). Training products of experts by minimizing contrastive divergence. Neural Computation, 14(8), 1771-1800.

[49] Hoffmann F., B. Baesens, C. Mues, T. Gestel y J. Vanthienen (2007). Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. European Computer Journal of Operational Research, 177(1), 540-555.

[50] Hsieh, N.C. (2005). Hybrid mining approach in the design of credit scoring models. Expert Systems with Applications, 28(4), 655-665.

[51] Huang, Z., H. Chen, C.J. Hsu, W.H. Chen y S. Wu (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. Decision Support Systems 37(4), 543-558.

[52] Huang, J., H. Tzeng y S. Ong (2006). Two stage genetic programming for the credit scoring. Computer Journal of Applied Mathematics and Computation, 14(3), 1039-1053.

[53] Huang, C.L., M.C. Chen y C.J. Wang (2007). Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 33(4), 847-856.

[54] Huysmans, J., B. Baesens, J. Vanthienen y T. Gestel (2006). Failure prediction with self organizing maps. Computer Journal of Expert Systems with Applicacions, 30(4), 479-487.

[55] Islam, M.J., Q.M.J. Wu, M. Ahmadi y M. Sid-Ahmed (2007). Investigating the performance of Naive-Bayes classifiers and K-nearest neighbor classifiers. International Conference on Convergence Information Technology. IEEE Computer Society.

[56] Jensen, L. (1992). Using neural networks for credit card accounts. Computer Journal of Managerial Finance, 18(15), 26-29.

[57] Johnson, R.W. (2004). Legal, social, and economic issues in implementing scoring in the united states. En Thomas, Edelman, and Crook (eds) Readings in Credit Scoring: Recent Developments, Advances, and Aims, 5-15. Oxford University Press.

[58] Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29(2), 119–127.

[59] Kumar, K., y S. Bhattacharya (2006). Artificial neural network vs linear discriminant analysis in credit ratings forecast: a comparative study of prediction performances. Review of Accounting and Finance, 5(3), 216-227.

[60] Laitinen, E.K. (1999). Predicting a corporate credit analyst's risk estimate by logistic and linear models. International Review of Financial Analysis 8(2), 97-121.

[61] Lane, S. (1972). Submarginal credit risk classification. Journal of Financial Quantitative Analysis, June, 313-328.

[62] Lee, T.S. y I.F. Chen (2002). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications, 28(4), 743-752.

[63] Lee, T.S., C.C. Chiu, C.J Lu y I.F. Chen (2002). Credit scoring using the hybrid neural discriminant technique. Expert Systems with Applications, 23(3), 245-254.

[64] Lee, T.S., C.C. Chiu, Y.C. Chou y C.J. Lu (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics & Data Analysis, 50, 1113-1130.

[65] Lewis, E.M. (1992). An Introduction to Credit Scoring, 2da. Ed. Athena Press, CA.

[66] Li, X.S. y Y.H. Guo (2006). Personal credit scoring models on Naive Bayesian classifier. Computer Engineering and Applications, 42(1), 197-201.

[67] Li, F. C. (2009). The hybrid credit scoring model based on KNN classifier. En Sixth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE Computer Society.

[68] Li, H., J. Sun, y J. Wu (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. Expert Systems with Applications, 37(8), 5895-5904.

[69] Li, T., W. Shiue y W. Huang (2006). The evaluation of consumer loans using neural networks. Omega Computer Journal, 31(2), 83-96.

[70] Makowski, P. (1985). Credit scoring branches out. Credit World, 75, 30-37.

[71] Malhotra R. y K. Malhotra (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. Computer Journal of Operational Research, 36(2), 190-201.

[72] Marinakis, Y., M. Marinaki, M. Doumpos, N. Matsatsinis y C. Zopounidis (2008). Optimization of nearest neighbor classifiers via metaheuristic algorithms for credit risk assessment. Journal of Global Optimization 42(2), 279-293.

[73] Martens, D., B. Baesens, T.V. Gestel y J. Vanthienen (2007). Comprehensive credit scoring models using rule extraction from support vector machines. European Journal of Operational Research, 183, 1466-1476.

[74] Martens, D., T.V. Gestel, D.B. Manu, R. Haesen, J. Vanthienen y B. Baesens (2008) Credit rating prediction using ant colony optimization. Journal of the Operational Research Society, 61, 561-573.

[75] Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society of London, 209, 415–446.

[76] Mileris R. y V. Boguslauskas (2010). Data reduction influence on the accuracy of credit risk estimation models. Inzinerine Ekonomika-Engineering Economics, 21(1), 5-11.

[77] Min, J.H. y Y.C. Lee (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Systems with Applications, 28, 603-614.

[78] Moody's (2000). RiskCalcTM for private companies: Moody's default model. Disponible en www.moodys.com

[79] Moses, D. y S.S. Liao (1987). On developing models for failure prediction. Journal of Commercual Bank Lending, 69, 27-38.

[80] Mues, C., B. Baesens, C.M. Files y J. Vanthienen (2004). Decision diagrams in machine learning: an empirical study on real-life credit-risk data. Expert Systems with Applications 27(2), 257-264.

[81] Myers, J.H. y E.W. Forgy (1963). The development of numerical credit evaluation systems. Journal of American Statistics Association, 58, 799-806.

[82] Narain B. (1992). Survival Analysis and the credit granting decision. In Credit Scoring and Credit Control, L.C.Thomas, J.N.Crook, D.B.Edelman (eds.), 109-122, Oxford University Press, Oxford.

[83] Nauck, D., F. Klawonn y R. Kruse (1997). Foundations of Neuro-Fuzzy Systems. Wiley, New York.

[84] Ong, C.S., J.J. Huang y G.H. Tzeng (2005). Building credit scoring models using genetic programming. Expert Systems with Applications, 29(2), 41-47.

[85] Ooghe H., H. Claus, N. Sierens N. y J. Camerlynck (1999). International comparison of failure prediction models from different countries: an empirical analysis. Working Paper, 99/79, Univesity of Ghen.

[86] Orgler, Y.E. (1970). A credit scoring model for commercial loans. Journal of Money, Credit and Banking, 2(4), 435-445.

[87] Orgler, Y.E. (1971). Evaluation of bank consumer loans with credit scoring models. Journal of Bank Research, 2(1), 31-37.

[88] Paredes, R., y E. Vidal (2000). A class-dependent weighted dissimilarity measure for nearest neighbor classification problems. Pattern Recognition Letters 21(12), 1027-1036.

[89] Pearl, J. (1988). Probabilistic reasoning in Intelligent Systems: networks for plausible inference. Morgan Kaufmann, San Francisco, CA.

[90] Piramuthu, S. (1999). Financial credit risk evaluation with neural and neuro fuzzy systema. Computer Journal of Operational Research, 112(6), 310-321.

[91] Platt, H.D. y M.D. Platt (1990). Development of a class of stable predictive variables: the case of bankruptcy prediction. Journal of Business Finance and Accounting, 17, 31-51.

[92] Quinlan, J.R. (1979). Discovering rules from large collections of examples: a case study. En D. Michie (Ed.), Expert Systems in the Micro-electronic Age, 68-201.

[93] Quinlan, J.R. (1986) Induction of decision trees. Machine Learning 1(1), 81-106.

[94] Quinlan, J.R. (1993). C4.5 programs for machine learning. Morgan Kaufmann, San Francisco, CA.

[95] Sarlija, N., M. Bensic y M.Z. Susac (2009). Comparison procedure of predicting the time to default in behavioural scoring. Expert Systems with Applications, 36(5), 8778–8788.

[96] Setiono, R., y H. Liu (1996). Symbolic representation of neural networks. IEE Computer, 29(3), 71-77.

[97] Sexton, R. S., S. McMurtrey y D.J. Cleavenger (2006). Knowledge discovery using a neural network simultaneous optimization algorithm on a real world classification problem. European Journal of Operational Research, 168, 1009-1018.

[98] Shin, K.S., T.S. Lee y H.J. Kim (2005). An application of support vector machines in bankruptcy prediction model. Expert Systems with Applications, 28(1), 127-135.

[99] Sohn, S.Y. y H.W. Shin (2006). Reject inference in credit operations based on survival analysis. Expert Systems with Applications, 31(1), 26–29.

[100] Steenackers, A. y M.J. Goovaerts (1989). A credit scoring model for personal loans. Insurance: Mathematics and Economics, 8(1), 31-34.

[101] Stepanova, M. y L.C. Thomas (2002). Survival analysis methods for personal loan data. Operations Research, 50(2), 277-289.

[102] Suykens, J.A.K. and J. Vandewalle (1999). Least squares support vector machine classifiers. Neural Process Letters, 9, 293-300.

[103] Suykens, J.A.K., T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle (2002). Least squares support machines. World Scientific Publishing Co., Pte, Ltd. Singapore.

[104] Taffler, R. J. y B. Abassi (1984). A model for predicting debt servicing problems in developing countries. Journal of the Royal Statistical Society, 147(4), 541-568.

[105] Tapiero, C.S. (2010). Risk finance and asset pricing: value, measurements, and markets. Wiley.

[106] Tang, T. C. y L.C. Chi (2005). Predicting multilateral trade credit risks: comparisons of logit and fuzzy logic models using ROC curve analysis. Expert Systems with Applications, 28(3), 547-556.

[107] Thomas L.C., D.B. Edelman y J.N. Crook (2002). Credit Scoring and its applications. Society for Industrial and Applied Mathematics, Philadelphia.

[108] Tsai C.F. y M.L. Chen (2010). Credit rating by hybrid machine learning techniques. Applied Soft Computing, 10, pp.374-380.

[109] Turney, P. (2000). Types of cost in inductive concept learning. Presentado en Workshop on Cost-Sensitive Learning en Seventeenth International Conference on Machine Learning, Stanford University, Stanford, CA.

[110] Vapnik, V. (1998). Statistical learning theory. John Wiley: New York.

[111] Wang, Y., S. Wang, S. y K.K. Lai (2005). A new fuzzy support vector machine to evaluate credit risk. IEEE Transactions on Fuzzy Systems, 13(6), 820-831.

[112] Wang, G., J. Hao, J. Ma y H. Jiang (2011). A comparative assessment of ensemble learning for credit scoring. Expert systems with applications, 38, 223-230.

[113] West, D. (2000). Neural network credit scoring models. Computers and Operations Research, 27(3), 1131-1152.

[114] West, D., S. Dellana y J. Qian (2005). Neural network ensemble strategies for financial decision applications. Computer Journal of Operations Research, 32(2), 2543-2559.

[115] Wiginton, J.C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. Journal of Quantitative Analysis, 15, 757-770.

[116] Witten I.H., E. Frank y A.H. Mark (2011). Data mining: practical machine learning tools and techniques. Elsevier, Morgan Kaufmann.

[117] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5, 241–259.

[118] Wolpert, D.H. (1996). The lack of a priori distinctions between learning algorithms. Neural Computation, 8, 1341-1390.

[119] Zhao, H. (2007). A multi-objective genetic programming approach to developing pareto optimal decision trees. Decision Support Systems, 43(3), 809-826.

[120] Zhu, H., P.A. Beiling y G.A. Overstreet (2001). A study in the combination of two consumer credit scores. Journal of the Operational Research Society, 52(9), 974-980.

Apéndice

Tablas descriptivas de los conjuntos de datos

Conjunto	Descripción	#Atributos (Num/Categ)	#Instancias (Malos/Buenos)	Problema
Alemán**	Información relativa a solicitantes de crédito proporcionada por Dr. Hans Hofmann	20 (7/13)	1,000 (300/700)	Estático: determinar el perfil de riesgo de nuevas solicitudes de crédito
Australiano **	Otorgamiento de tarjetas de crédito proporcionado por Quinlan (1986) omitiendo nombres y descripción de los atributos	14 (6/8)	690 (307/383)	Estático: Clasificación buenos y malos solicitantes
PAKDD***	Información utilizada en el reto PAKDD 2009, la cual corresponde solicitudes de tarjetas de crédito aceptadas. El conjunto de entrenamiento y prueba son no adyacentes en el tiempo.	22 (10/12)	50,000 (9874/40126)	Estático: Clasificación buenos y malos clientes
Cerveza	Registros sobre venta de cerveza durante 2009 y 2010 de cierta empresa mexicana. Se otorgaron líneas de crédito a algunos clientes.	16 (2/14)	3,200 de los cuales 2,156 con línea de crédito (949/1207)	Estático: Clasificación de clientes con y sin línea de crédito así como del perfil de riesgo de los acreditados
Autos	Información de créditos de auto otorgados entre 2006 y 2009 empresa mexicana especializada en el financiamiento automotriz.	24 (17/7)	48,550 distintos créditos	Estático: Clasificación buenos/malos Dinámico*: Clasificación de clientes con distintos perfiles de riesgo bajo distintas dimensiones de análisis: fecha de originación, morosidad en pagos vencidos (ventanas de observación móviles)

Tabla 1. Resumen de los cinco conjuntos de datos disponibles

Atributo	Descripción	Medición
Cheques	Saldo cuenta actual en marcos alemanes (DM)	Categórica; 4 valores: A11-Saldo a favor, A12-[0,200), A13-Mayor o igual a 200, A14-Sin cuenta de cheques
Plazo	Duración en meses	Entero
Historial	Historial crediticio	Categórica; 5 valores: A30-Sin créditos/Al corriente, A31-Al corriente en este banco, A32-Al corriente en créditos actuales, A33-Pagos retrazados en el pasado, A34-Cuenta crítica/otros créditos en otros bancos
Proposito	Propósito del crédito	Nominal; 11 valores: A40-Auto nuevo, A41-Auto usado, A42-Muebles/Equipo, A43-Radio/Televisión, A44-Electrodomésticos, A45-Reparaciones, A46-Educacion, A47-Vacaciones, A48-Reentrenamiento, A49-Negocios, A410-Otros
Monto	Monto del crédito	Intervalo
Ahorros	Saldo en cuentas o inversiones, en DM	Categórica; 5 valores: A61-[0,100), A62-[100,500), A63-[500,1000), A64-Mayor o igual a 1000, A65-Desconocido/Sin cuenta de ahorros
TiempoEmpleo	Antigüedad en el empleo actual en años	Categórica; 5 valores: A71-Desempleado, A72-(0,1), A73-[1,4), A74-[4,7), A75-Mayor o igual a 7
TasaPago	Tasa de pago como porcentaje del ingreso disponible	Intervalo
Personal	Estatus civil y sexo	Nominal; 5 valores: A91-hombre divorciado/separado, A92-mujer divorciada/separada/casada, A93-hombre soltero, A94-hombre casado/viudo, A95-mujer soltera
Deudores	Otros deudores/garantes	Nominal; 3 valores: A101-Ninguno, A102-Co solicitante, A103-Garante
TiempoResidencia	Tiempo en la residencia actual	Intervalo
Propiedad	Tipo de propiedad	Nominal; 4 valores: A121-Bienes raíces, A122-Ahorro vivienda, A123-Auto u otro no considerado en atributo 'Ahorros', A124-Desconocido/Sin propiedad
Edad	Edad en años	Entero
OtrosBienes	Otros planes de inversión	Nominal; 3 valores: A141-Banco, A142-Tienda, A143-Ninguno
Vivienda	Tipo de vivienda	Nominal; 3 valores: A151-Rentada, A152-Propia, A153-Gratis
NumCreditos	Número de créditos en el banco actual	Entero
Empleo	Tipo de empleo	Categórica; 4 valores: A171-Desempleado/No residente sin capacitación,

		A172-Residente no capacitado, A173-Empleado Capacitado, A174-Administrador/Autoempleado/Altamente calificado/Director
Dependientes	Número de dependientes económicos	Entero
Telefono	Estatus de número telefónico	Binaria: A191-No tiene, A192-Sí tiene y está registrado a su nombre
Extranjero	Trabajador extranjero	Binaria: A201-Sí, A202-No
Clase	Indicadora del perfil crediticio	Binaria: 0-Bueno, 1-Malo

Tabla 2. Atributos del conjunto Alemán

Atributo	**Medición**	**Tasa Faltantes**
A1	Categórica; 2 valores: 0-a, 1-b	1.7%
A2	Intervalo	1.7%
A3	Intervalo	-
A4	Categórica; 3 valores: 1-p, 2-g, 3-gg	0.9%
A5	Categórica; 14 valores: 1-ff, 2-d, 3-i, 4-k, 5-j, 6-aa, 7-m, 8-c, 9-w, 10-e, 11-q, 12-r, 13-cc, 14-x	0.9%
A6	Categórica; 9 valores: 1-ff, 2-dd, 3-j, 4-bb, 5-v, 6-n, 7-o, 8-h, 9-z	1.3%
A7	Intervalo	1.3%
A8	Categórica; 2 valores: 0-t, 1-f	-
A9	Categórica; 2 valores: 0-t, 1-f	-
A10	Intervalo	-
A11	Categórica; 2 valores: 0-t, 1-f	-
A12	Categórica; 3 valores: 1-s, 2-g, 3-p	-
A13	Intervalo	1.9%
A14	Intervalo	-
A15	Binaria: 0-Bueno('-'), 1-Malo('+')	-

Tabla 3. Atributos conjunto de datos Australiano

Atributo	Descripción	Medición	Tasa Faltantes	
			Entrenamiento	Prueba
ID_SHOP	Código de la tienda donde se realizó la solicitud	Entero	-	-
SEX	Sexo	Binaria: M=Masculino, F=Femenino	0.006%	-
MARITAL_STATUS	Estatus civil	Nominal; 5 valores: S=Soltero, C=Casado, D=Divorciado, V=Viudo, O=Otro	-	-
AGE	Edad del aplicante	Entero	-	-
QUANT_DEPENDANTS	Número de dependientes económicos	Entero	100%	-
EDUCATION	Nivel de educación del solicitante	Nominal; 4 valores: 1, 2, 3, 4 (categorías no especificadas)	100%	10.51%
FLAG_RESIDENCIAL_PHONE	Cuenta con teléfono residencial	Binaria: Y=Si, N=No	-	-
AREA_CODE_RESIDENCIAL_PHONE	Código de área del teléfono residencial	Entero	-	-
PAYMENT_DAY	Día (fijo) de pago en el mes	Entero	-	-
SHOP_RANK	Calificación que la compañía asigna a la tienda	Nominal; 3 valores entrenamiento: 0, 2, 3 y 4 valores prueba: 0, A, B, C, D	-	-
RESIDENCE_TYPE	Tipo de residencia	Nominal; 4 valores: P=Propia, A=Rentada, C=DE los padres, O=Otra	-	0.36%
MONTHS_IN_RESIDENCE	Tiempo en residencia actual en meses	Entero	-	-
FLAG_MOTHERS_NAME	Se proporciona el nombre de la madre del solicitante	Binaria: Y=Si, N=No	-	-
FLAG_FATHERS_NAME	Se proporciona el nombre del padre	Binaria: Y=Si, N=No	-	-
FLAG_RESIDENCE_TOWN=WORKING_TOWN	Trabaja en el mismo pueblo donde vive	Binaria: Y=Si, N=No	-	-
FLAG_RESIDENCE_STATE=WORKING_STATE	Trabaja en el mismo estado donde vive	Binaria: Y=Si, N=No	-	-
MONTHS_IN_THE_JOB	Tiempo en empleo actual en meses	Entero	-	-
PROFESSION_CODE	Código de la profesión del aplicante	Entero	-	-
MATE_INCOME	Ingreso neto mensual de la pareja del aplicante en moneda brasileña (R$)	Intervalo	-	-
FLAG_RESIDENCIAL_ADDRESS=POSTAL_ADDRESS	Recibe el correo postal en el domicilio donde vive	Binaria: Y=Si, N=No	-	-
FLAG_OTHER_CARD	Si cuenta con otro crédito o	Binaria: Y=Si, N=No	100%	-

	tarjeta de marca propia			
QUANT_BANKING_ACCOUNTS	Número de cuentas bancarias del solicitante	Entero	100%	0.36%
PERSONAL_REFERENCE_#1	Nombre de la referencia personal #1 (en portugués)	Cadena	0.004%	1.45%
PERSONAL_REFERENCE_#2	Nombre de la referencia personal #2 (en portugués)	Cadena	10.40%	1.45%
FLAG_MOBILE_PHONE	Posee teléfono cellular	Binaria: Y=Si, N=No	100%	100%
FLAG_CONTACT_PHONE	Posee teléfono de contacto	Binaria: Y=Si, N=No	100%	-
PERSONAL_NET_INCOME	Ingreso neto mensual del solicitante en moneda brasileña (R$)	Intervalo	-	-
COD_APPLICATION_BOOTH	Código de módulo donde se realizó la solicitud	Entero	100%	-
QUANT_ADDITIONAL_CARDS_IN_THE_APPLICATION	Número de tarjetas adicionales solicitadas	Entero	-	100%
FLAG_CARD_INSURANCE_OPTION	Solicita seguro de tarjeta	Nominal igual a N=No	100%	-
TARGET_LABEL_BAD=1	Perfil de riesgo	Binaria: 1-Malo, 0-Bueno	-	-

Tabla 4. Atributos conjunto de datos PAKDD

Atributo	**Descripción**	**Medición**	**Tasa Faltantes**
Clave	Cliente clasificado como clave (por alto volumen de ventas)	Binaria: 1-Clave, 0-No Clave	7.8%
Imagen	Cliente clasificado como imagen (por ubicación o tipo de sector que atiende)	Binaria: 1-Cliente Imagen, 0-No Imagen	13.6%
Exclusivo	Estatus de exclusividad de la marca en los productos de cerveza que el cliente ofrece	Binario: 1-Exclusivo, 0-Mixto	-
Esquina	Ubicación del establecimiento	Binaria: 1-Esquina, 0-Otra ubicación	13.5%
Calle	Tipo de calle	Nominal; 3 valores: 1-Principal, 2-Secundaria, 3-Carretera	13.5%
Zona	Tipo de zona	Nominal; 4 valores: 1-Comercial, 2-Habitacional, 3-Industrial, 4-Mixta	13.5%
Trafico	Nivel de tráfico vehicular a que se expone el establecimiento	Categórica; 3 valores: 1-Alto, 2-Medio, 3-Bajo	-
Dimension	Dimensión de la fachada del establecimiento en metros	Intervalo	-
Local	Tipo de local	Nominal; 3 valores: 1-Empresa, 2-Propio, 3-Rentado	13.5%
Region	Estado de la república donde se ubica el establecimiento	Nominal; 3 valores: 1-Chiapas, 2-Campeche, 3-Tabasco	-

Giro	Giro del establecimiento	Nominal; 8 valores: 1-Abarrotes, 2-Cantina, 3-Cerveceria, 4-Depósito, 5-Evento aire libre, 6-Minisuper, 7-Restaurante, 8-Otros	-
Control	Grado de control o influencia en el establecimiento	Categórica; 5 valores: 1-Blindado, 2-Alto, 3-Mediano, 4-Bajo, 5-Sin control	0.3%
Tamaño	Tamaño del cliente (en función de sus ventas)	Categórica; 6 valores: 1-Exitoso, 2-Productivo, 3-Promedio, 4-Cobertura, 5-Apoyo, 6-Minimo	12.6%
NivelSocEcon	Nivel socioeconómico de la zona donde se ubica el establecimiento	Categórica; 3 valores: 1-Nivel AB, 2-Nivel C, 3-Nivel DE	-
Resultado	Ventas totales realizadas al cliente (kilolitros)	Intervalo	-
GEC	Tipo de cliente en función del promedio de ventas	Categórica; 5 valores: 1-Platino, 2-Oro, 3-Plata, 4-Bronce, 5-Customizado	-
SVencido90	Indicadora de incumplimiento (saldo vencido 90+ días)	Binario: 1-Incumplido, 0-No incumplido	-
LineaCredito	Indicadora de otorgamiento de línea de crédito al cliente	Binario: 1-Con crédito, 0-Sin crédito	-

Tabla 5. Atributos conjunto de datos Cerveza

	Distribución								
Atributo	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Clave	76%	24%							
Imagen	94%	6%							
Exclusivo	21%	79%							
Esquina	84%	16%							
Calle		23%	33%	44%					
Zona		15%	79%	1%	5%				
Trafico		31%	27%	42%					
Local		2%	61%	37%					
Region		26%	1%	74%					
Giro		17%	4%	14%	25%	4%	9%	12%	15%
Control		16%	7%	5%	21%	50%			
Tamaño		5%	16%	27%	16%	37%			
NivelSocEcon		2%	77%	21%					
GEC		7%	34%	37%	12%	11%			
LineaCredito	33%	67%							
SaldoVencido90[44]	56%	44%							

Tabla 6. Distribución de los atributos categóricos conjunto Cerveza

[44] Por definición, las proporciones de clientes con saldo vencido son calculadas sobre el universo de clientes a quienes se les otorgó línea de crédito.

	Dimensión	Resultado
Min	1.00	0.04
Max	650.00	29,169.91
Promedio	11.09	496.86
Desv Est	20.38	1,142.17

Tabla 7. Estadísticas descriptivas de los atributos numéricos conjunto Cerveza

Atributo	Descripción	Medición	Tasa Faltantes
educacionnueva	Máximo nivel de estudios	Nominal; 9 valores: 21-Primaria, 22-Secundaria, 23-Preparatoria, 24-Carrera, 25-Universidad, 26-Diplomados, 27-Posgrado, 28-Maestria, 29-Doctorado	
EDOCIVIL	Estado civil	Nominal; 5 valores: 0-Soltero, 1-Divorciado, 2-Viudo, 3-Unión libre, 4-Casado	
SEXO	Género	Nominal; 2 valores: 0-Masculino, 1-Femenino	
EDAD	Años cumplidos al solicitar el crédito	Entero	
DEPENDIENTES	Número de personas que dependen del solicitante	Intervalo	
INGRESOS	Monto de ingresos mensuales	Intervalo	0.7%
HIPOTECA	Cuenta con crédito hipotecario	Binario: 0-No tiene, 1-Si tiene	
INGRESOSNETOS	Monto de ingresos netos mensuales	Intervalo	0.7%
OTROSINGRESOS	Otros ingresos (mensuales)	Intervalo	0.2%
DEUDAS	Monto de obligaciones (deudas) actuales	Intervalo	0.1%
DOMICILIACION	Cuenta con domiciliación de pagos	Binario: 0-No tiene, 1-Si tiene	
POP	Probabilidad de pago al solicitar el crédito	Intevalo	
SCORE	Score calculado al solicitar crédito	Intervalo	
plazo	Plazo del crédito en meses	Intervalo	
montofin	Monto a financiar	Intervalo	
mensualidad	Monto de la mensualidad	Intervalo	
Precio	Precio del vehículo	Intervalo	
Enganchep	Tasa de enganche (%)	Intervalo	
TasaSTD	Tasa teórica a la que le prestaría la institución (%)	Intervalo	
TasaCliente	Tasa a la que la institución prestó al	Intervalo	

	cliente (%)		
Marca	Marca del automóvil a financiar	Nominal; 8 valores: 1-FORD, 2-CHEVRO, 3-LAND R, 4-LAND ROVER, 5-LINCOLN, 6-MAZDA, 7-SEAT, 8-VOLVO	
Año	Año de inicio del contrato	Intervalo; en formato aaaa	
Mes	Mes de inicio del contrato	Intervalo; en formato mm	
subsidiado	Se otorgó subsidio en el préstamo	Binaria: 0-Ni tiene, 1-Si tiene	

Tabla 8. Atributos conjunto de datos Autos (Base socio-demográfica)

Atributo	Descripción
año	Año del reporte
Mes	Mes del reporte
Dpd	Días de saldo vencido
Pagadas	Mensualidades pagadas
Remterm	Mensualidades faltantes
saldoinsoluto	Saldo insoluto
Perdida	Indicadora de incumplimiento (+90días de mora)

Tabla 9. Atributos conjunto datos Autos (Base comportamiento)

Alemán	Australiano	PAKDD	Cerveza	Autos
A1	A8	MARITAL_STATUS	Control	INGRESOSNETOS
A3	A10	AGE	Cliente	DEUDAS
A2	A9	FLAG_RESIDENCIAL_PHONE	Giro	EDAD
A5	A14	AREA_CODE_RESIDENCIAL_PHONE	Region	Montofin
A4	A7	MONTHS_IN_THE_JOB	GEC	DEPENDIENTES
A9	A5	PROFESSION_CODE	Calle	HIPOTECA
A21	A6	MATE_INCOME	Exclusivo	Enganchep
A10	A3	QUANT_DEPENDANTS	Tamaño	OTROSINGRESOS
A6	A13	MONTHS_IN_RESIDENCE	Trafico	ESTADOCIVIL

Tabla 11. Atributos con mayor contribución estable en la explicación de la variable de clase

Citas

[5] Cita de un desplegado en mayo de 2004 sobre la plática de Alan Greenspan, presidente de la Reserva Federal de EUA que dio en octubre de 2002 a la American Bankers Association: http://www.federalreserve.gov

En el contexto nacional, conviene citar el siguiente extracto de la sección de exposición de motivos de la Ley de Instituciones de Crédito (Publicada en el Diario Oficial de la Federación el 18 de julio de 1990. Actualizada con las modificaciones del Decreto publicado en el Diario Oficial de la Federación el 25 de mayo de 2010, http://www.cnbv.gob.mx/Bancos/Paginas/Normatividad.aspx):

"...Así las cosas, es obligación de esta Soberanía sentar las bases jurídicas apropiadas para que, por una parte, la oferta de créditos quede al alcance de aquéllos que aún no han podido disfrutar de sus beneficios y, por la otra, la ciudadanía tenga la confianza de aceptarlos.

Cabe destacar que el crédito representa una oportunidad para que las empresas desarrollen actividades productivas y comerciales y para que las personas puedan adquirir bienes y

servicios para su consumo. Por ello, el otorgamiento de crédito es una de las condiciones necesarias para fomentar el crecimiento económico del país y aumentar el nivel de vida de la población.

Los beneficios del crédito no se reducen a la relación entre acreditante y acreditado, sino que se extienden a toda la sociedad. El acceso al financiamiento facilita la creación de nuevas empresas y negocios, así como la expansión de los ya existentes, lo cual, a su vez, se refleja en el aumento de la actividad económica, la creación de empleos y la demanda de productos y servicios. En este sentido, la reactivación del crédito puede ser el detonante de un círculo virtuoso para el desarrollo nacional..." [publicado en el Diario Oficial de la Federación el 30 de noviembre de 2005]