Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Físico Matemáticas

Modelos de score de crédito aplicados a instituciones bancarias

Como requisito parcial para la obtención del grado de

Licenciado en Actuaría

Por

Salvador Alejandro Uribe Castellanos

Asesorado por

M. en F. Jorge Luis Reyes García

Puebla, Puebla

Febrero 2021

Agradecimientos

Reconocer a los colaboradores de la Facultad de Ciencias Físico Matemático, en particular a los profesores de la licenciatura en actuaria por brindarme los conocimientos necesarios para ejercer la carrera de mis sueños.

Doy gracias a todas las personas que me apoyaron e hicieron posible que este trabajo se realice con éxito, especialmente a mi madre Carmen C. que me ha enseñado lo que es el amor verdadero, mi padre Rafael U. que es mi ejemplo a seguir, mi hermano Rafael U. que nunca me ha dejado sólo y mi mentor Ernesto quien nunca dejó de confiar en mi y siempre me motivó cuando más lo necesitaba.

Agradezco a todas las personas que me acompañaron a lo largo de mi estadía en la universidad, especialmente a Marypaz N., Ana F. y Laura C., quienes me apoyaron todos los días a seguir adelante, no sólo como estudiante, sino también como persona de igual manera a Lilia M. quien me motivó a realizar la tesis.

También quiero darle las gracias a mi profesor y asesor M. en F. Jorge Luis Reyes García por el tiempo dedicado y los conocimientos brindados. A mis sinodales Carlo Ezra Martínez Ruiz, Brenda Zavala López y Víctor Hugo Vázquez Guevara por tomarse el tiempo necesario para revisar este trabajo y orientarme.

Sin duda alguna, este trabajo ha sido uno de los proyectos más interesantes y largos que he realizado hasta hoy, tuve momentos difíciles, pero sin importar el obstáculo el proyecto salió a flote, mi recomendación es que la clave del éxito se encuentra en la perseverancia y consistencia.

Contenido

Introducción

Justificación del trabajo e importancia con la práctica actuarial

En las instituciones financieras un peligro común es el riesgo crediticio: “Se puede definir como la pérdida potencial producto del incumplimiento de la contraparte en una operación que incluye un compromiso de pago” de Lara (2008). Una manera de reducir este riesgo es por medio de los modelos de calificación crediticia, los cuales Araujo & Carmona (2007) los describió de la siguiente manera “Los modelos de calificación crediticia son sistemas que asignan calificaciones a las variables de decisión de crédito mediante la aplicación de técnicas estadísticas. Estos modelos tienen como objetivo identificar características que pueden diferenciar entre buenos y malos créditos”.
Dichos modelos sirven para aceptar o rechazar una solicitud de préstamo y para ajustar la tasa de interés que se le aplicará, la cual va de la mano con el riesgo que tiene esta inversión ya que con un mayor riesgo se espera una ganancia más alta, por lo tanto, es complicado saber si conviene realizar la inversión si no se sabe el nivel de riesgo que conlleva.

El propósito principal de este trabajo es encontrar el mejor modelo enfocado a un conjunto de datos para reducir las pérdidas generadas debido a deudas incobrables por clientes a quienes no se les debía haber aceptado la solicitud. Una ventaja de usar los modelos de calificación es que, si ciertos patrones cambian en una población, el modelo realiza los ajustes pertinentes por medio del historial crediticio que se genera con el tiempo y de esta manera no tener pérdidas provocadas por una falta de actualización.

El modelo de calificación crediticia abarca conocimientos profundos y de diferentes áreas como finanzas, estadística y programación, por ello un profesional de la actuaría es más que pertinente para enfrentar este tipo de problemas.

Objetivos de la tesis

El objetivo de esta tesis es analizar diferentes métodos estadísticos que existen para la modelación de calificación crediticia, el cual estará enfocado a un conjunto de datos realizados en el 2005 por medio de una institución financiera en Taiwán el cual contiene 30,000 registros con 24 atributos. Este conjunto de datos contiene variables cuantitativas y cualitativas que describen a las personas que solicitaron un préstamo.
En la actualidad uno de los principales modelos que se usa en las instituciones para la clasificación crediticia es regresión logística y en este trabajo se evaluó su desempeño y se analizó la principal razón por la cual este modelo tiene tanta fama hoy en día.

Metodología de la investigación

Los métodos estadísticos aplicados en este trabajo son de aprendizaje supervisado, haciendo referencia a introducir las categorías (modelo de regresión logística, análisis discriminante lineal, árbol de decisión, modelo de redes neuronales).

Una vez aplicado cada método analizado al conjunto de datos, se realizaron diversas pruebas a cada clasificador, con el fin de tener métricas que nos ayuden a contrastar cada modelo directamente, por último, se realizó una comparación entre todas las pruebas y se analizarán las ventajas y desventajas que tiene un modelo con otro y de esta manera, poder concluir el mejor modelo posible para este conjunto de datos.

Capítulo 1.- Riesgo de crédito

En este capítulo se comentan las actividades más relevantes de las empresas que laboran en el sector financiero y el riesgo que conlleva al realizar cada una de estas. Se analizó la definición y clasificación de riesgo, ya que se puede presentar en más de una manera. Se resaltó el riesgo de crédito por ser inherente a las actividades de las instituciones financieras y ser el principal motivo por el cual se realizaron los diferentes modelos matemáticos en este trabajo. Por último se comentó el proceso que deben de realizar las empresas para reducir el riesgo, llamado administración de riesgos, de la misma manera se presenta un histórico de los desastres financieros más relevantes por la falta de este proceso.

1.1 Principales actividades del sector bancario

Como se analizó en la justificación del trabajo, tenemos que el riesgo de crédito es la pérdida potencial que puede sufrir una institución ante el incumplimiento de pago de dinero de la contraparte, sabiendo esto es fácil ver que las instituciones en donde se practican los créditos sufren en gran medida este riesgo y un claro ejemplo es el banco. Cabe mencionar que en este capítulo se entiende como banco a la banca múltiple.

Para poder entender de mejor manera la forma en que el banco sufre este riesgo, es relevante saber la función principal que realizan estas instituciones y según la Comisión Nacional Bancaria y de Valores [CNBV] es:

> La función de banca y crédito que las instituciones de banca múltiple proporcionan a sus clientes consiste en captar los recursos dispersos en la economía, conjuntarlos en ahorro y canalizarlos en forma de financiamiento (créditos) hacia individuos o empresas que generen valor agregado en la economía. Por ello, contar con un sistema bancario fuerte y eficiente fomenta el crecimiento económico del país.

Con la definición anterior nos damos cuenta de que la actividad de realizar créditos es de suma importancia para el banco y este es un pilar para el crecimiento económico del país, por lo tanto, si el sistema bancario no tuviera una sólida protección ante el riesgo de crédito, no solo afectaría a la misma institución, también al país en donde se encuentra.

Un ejemplo claro de la relevancia que tiene el banco ante la sociedad lo comentó Cuartas (2013):

> Una entidad, empresa o persona que necesite dinero prestado, o quisiera negociar valores, no fácilmente conocería, quién o quiénes tienen dinero, y estarían dispuestas a prestarlo o están en condiciones de negociar confiablemente instrumentos negociables, tendría que averiguar quiénes disponen de dinero o títulos, con la gran dificultad de adquirir confianza, para negociar. Sin la existencia de la banca se estaría atado para negociar.

Con el ejemplo anterior, nos percatamos de que si no existe algún intermediario entre oferentes y demandantes, puede llegar a ser complicado encontrar a una persona física o empresa que tenga dinero suficiente para dar un crédito, y aún más complicado que también esa misma persona tenga la confianza de darlo, sin importar que tengas un buen perfil crediticio o no, ya que, quien dará el crédito no podrá respaldar tu credibilidad si no te conoce, por lo tanto el banco cumple con la tarea de facilitar este proceso.

1.2 Crédito

Se sabe que el crédito es un factor básico en la economía de un país y como se argumentó en la sección anterior tenemos que para el banco es una de las operaciones más importantes para su función, pero acaso ¿sabemos qué es el crédito? explicó Morales (2015) sobre la definición de crédito:

> El crédito es un préstamo en dinero, donde la persona se compromete a devolver la cantidad solicitada en el tiempo o plazo definido según las condiciones establecidas para dicho préstamo, más los intereses devengados, seguros y costos asociados si los hubiere.

Como lo menciona el autor en el párrafo anterior existen algunos créditos con costos asociados, por ejemplo, si al crédito se le agrega un seguro, el cual sirve por si el cliente fallece y no termina de pagar el crédito, el seguro cubre estas pérdidas por parte de la institución financiera.

Los intereses devengados hacen referencia al dinero extra que cobra la institución financiera o persona al cliente por haber prestado su dinero y también haber estado expuesto a que la otra persona no cumpliera con las condiciones establecidas, los intereses a cobrar varían según el riesgo que conlleva dicha operación y el periodo de tiempo que tardará el cliente en pagar el crédito.

> "La espera, que conlleva una pérdida de oportunidad y riesgo para el acreedor, explica la existencia del pago de intereses en las operaciones crediticias" (Morales, 2015).

El termino crédito según la Real Academia Española [RAE] (2020), "proviene del latín *creditum,* de *credere*, tener confianza", esto da a entender que la confianza es la base del crédito, aunque al mismo tiempo implica un riesgo, el crédito sin la confianza no existe, por lo tanto, el crédito es confianza.

De la misma manera, argumentó Jacques (2009) sobre el crédito: "Todo crédito implica una cierta confianza, de parte del que lo concede, en el beneficiario de éste. En otras palabras, la persona que recibe el crédito debe gozar de credibilidad". La principal función de que exista credibilidad en la persona que solicita el crédito es para tener una mayor certeza de que el cliente cumplirá con lo acordado, por lo tanto, el riesgo asumido es menor.

Del mismo modo Leroy (1992) afirmó que en la mayoría de las instituciones financieras para poder solicitar un crédito existen requisitos: "dichos requisitos son: identidad del prestatario, cantidad de dinero reembolsable al vencimiento del instrumento, cantidad correspondiente a intereses y fecha de pago". Estos requisitos pueden variar dependiendo la institución y el tipo de crédito que se desee, pero en particular se convierten en la credibilidad que el cliente presenta frente a la institución, si brinda una mayor credibilidad con un mejor ingreso mensual o un trabajo formal, tendrá más oportunidad de que le concedan el crédito.

Por lo tanto, concluimos con los argumentos de los autores anteriores, que la credibilidad es clave en el otorgamiento de créditos, el cual es otro punto a favor de los modelos estadísticos

para la aceptación o rechazo de solicitudes, dado que la probabilidad que genere el solicitante por medio del modelo será interpretado como la credibilidad que tiene este mismo, si tiene una probabilidad alta de pagar, su credibilidad será alta, en cambio sí tiene una probabilidad baja, su credibilidad ante la institución será baja.

¿Por qué realizar un crédito? Zorrilla (1994) Argumentó cuales han sido las 3 principales ventajas de realizar un crédito:

1.- Poner capital a disposición de quien no lo posee y facilita la disponibilidad de capitales a las personas que tienen aptitudes para utilizarlos.
2.- Facilitar el uso del pequeño ahorro, con la acumulación de los pequeños ahorros se forman grandes capitales para ser aplicados a la creación de empresas importantes.
3.- Ahorrar el uso de la moneda y en esa forma da mayor elasticidad y volumen a las operaciones de comercio.

En algunas ocasiones las instituciones o personas tienen una falta de liquidez que ocasiona estragos en su economía, para tener liquidez de dinero en un periodo de tiempo corto es común vender bienes a precios inusuales, o rechazar ofertas de inversión de muy buena calidad, lo que genera pérdidas notables, una manera de solucionar este tipo de casos es solicitando un crédito.

Una vez analizada la definición de crédito y las razones por las cuales es útil solicitarlo, debemos conocer los aspectos que tradicionalmente las instituciones financieras han tomado en cuenta para otorgar un crédito. Una manera fácil de conocer dichos aspectos es por medio del modelo “Cinco C´s del crédito: capacidad, capital, colateral, carácter y convivencia” (Gómez, 2014).

Capacidad. La capacidad de repago está relacionada con la volatilidad de los ingresos, ya que, a mayor volatilidad, mayor probabilidad de que aparezcan problemas a la hora de satisfacer los pagos de la deuda (Valle, 2015).

Como se argumentó en la definición anterior, las instituciones se inclinan por los clientes que tienen un ingreso más estable por medio de un trabajo formal, ya que gracias a ello es muy probable que tengan el dinero suficiente para pagar, mientras que por otro lado si el cliente tiene un ingreso volátil, es probable que el cliente incumpla con los pagos por no haber recibido el ingreso que él esperaba.

Capital. El capital hace referencia a varios ratios financieros, tales como su grado de apalancamiento (ratio deuda/capital propio) o su capacidad de servicio de la deuda (BAII/Intereses). Alto apalancamiento y escasa capacidad de cobertura del pago de intereses suelen estar asociados con altas probabilidades de fallido (Valle, 2015).

De la definición anterior tenemos que, si el cliente está endeudado y los pagos que realiza con respecto de su ingreso mensual es considerable alto, entonces tenemos que este cliente tiene probabilidad alta de que no pueda pagar el crédito, por lo tanto, se prefiere que las deudas del cliente no sean significativas con respecto de su ingreso.

Colateral. Para otorgar un crédito, suele requerirse la entrega de una garantía "colateral", en forma de bienes muebles o inmuebles, como inventarios o edificios, que serán aplicados para hacer frente a las obligaciones contraídas por el solicitante, en caso de que éste no pueda hacerlo por medios propios. Otro de los colaterales a los que con más frecuencia se acude en México, es el otorgamiento de avales personales por parte de personas de reconocida solvencia moral y material (Gómez, 2014).

Cabe recalcar que cuando existe una garantía colateral, el riesgo que asume el banco es menor, dado que, si el cliente deja de pagar, la garantía pasa a ser propiedad de la institución.

"Carácter. El carácter hace referencia a la reputación de la empresa en su sector, su antigüedad y la solidez percibida de sus operaciones" (Valle, 2015).

El carácter hace referencia a la credibilidad que tiene la empresa ante la sociedad, cabe recalcar que era esperado encontrarse con este aspecto por lo mencionado al principio de la sección.

Conveniencia. Se refiere a que tanto el deudor, como el intermediario, deben estar en posibilidades de obtener un rendimiento adecuado de los créditos otorgados. En la medida en que el margen de intermediación (i.e. la diferencia entre intereses cobrados y pagados por el intermediario) y la probabilidad de recuperación del crédito sean más elevadas, le convendrá más al intermediario otorgarlo (Gómez, 2014).

Como en toda inversión, existe una regla que es "a mayor riesgo mayor rendimiento", por lo tanto, la tasa de interés que estipule la institución debe de ser a un nivel tal que, le resulte beneficiosa

tomando en cuenta el riesgo en el que incurre, mientras que por otro lado el cliente debe de considerar si también para él le es beneficioso obtener el crédito con esa tasa de interés.

1.2.1 Tipos de crédito

Ahora que ya entendemos la definición de crédito, se presenta a continuación los tipos de crédito según la Comisión Nacional Bancaria y de Valores:

> **De Consumo**: a los créditos directos, incluyendo los de liquidez que no cuenten con garantía de inmuebles, denominados en moneda nacional, extranjera, en UDIs, o en VSM, así como los intereses que generen, otorgados a personas físicas, derivados de operaciones de tarjeta de crédito, de créditos personales, de créditos para la adquisición de bienes de consumo duradero (conocidos como ABCD), que contempla entre otros al crédito automotriz y las operaciones de arrendamiento financiero que sean celebradas con personas físicas; incluyendo aquellos créditos otorgados para tales efectos a los exempleados de las Instituciones (CNBV, 2016).

Este tipo de crédito es muy importante ya que es de los más utilizados y se comprende que generalmente los pagos son mensuales de acuerdo con el plan, la capacidad de pago, el monto y la institución.

> **Hipotecaria de Vivienda**: a los créditos directos denominados en moneda nacional, extranjera, en UDIs, o en VSM, así como los intereses que generen, otorgados a personas físicas y destinados a la adquisición, construcción, remodelación o mejoramiento de la vivienda sin propósito de especulación comercial; incluyendo aquellos créditos de liquidez garantizados por la vivienda del acreditado y los otorgados para tales efectos a los ex-empleados de las Instituciones (CNBV, 2016).

Cabe mencionar que este tipo de crédito no es fácil que lo acepten porque el monto suele ser grande, así que el cliente debe de tener un historial crediticio lo suficientemente bueno para que la institución pueda tener la confianza de que el crédito será pagado. Lo más importante sobre este crédito es que, dado que el dinero a prestar es relevante está respaldado por una garantía

hipotecaria. Es decir, si el deudor del crédito no pudiera pagar las cuotas, el acreedor podría llegar a quedarse con el activo hipotecado, o sea, con la casa que el deudor intentó comprar desde el principio.

> **Comercial**: a los créditos directos o contingentes, incluyendo créditos puente denominados en moneda nacional, extranjera, en UDIs, o en VSM, así como los intereses que generen, otorgados a personas morales o personas físicas con actividad empresarial y destinados a su giro comercial o financiero; incluyendo los otorgados a entidades financieras distintos de los de préstamos interbancarios menores a 3 días hábiles; las operaciones de factoraje y operaciones de arrendamiento financiero que sean celebradas con dichas personas morales o físicas; los créditos otorgados a fiduciarios que actúen al amparo de fideicomisos y los esquemas de crédito comúnmente conocidos como “estructurados”. Asimismo, quedarán comprendidos los créditos concedidos a entidades federativas, municipios y sus organismos descentralizados, cuando sean objeto de calificación de conformidad con las disposiciones aplicables (CNBV, 2016).

El crédito comercial se reduce a un aplazamiento del pago que una empresa concede a sus clientes o sea una facilidad de pago. Algunos créditos que se comprenden como comercial son los créditos puente o de factoraje, el primero facilita a la construcción de viviendas y el segundo a las transacciones derivadas de un contrato, donde una empresa vende sus cuentas por cobrar a una compañía financiera para que la empresa emisora pueda recibir su dinero más rápido.

1.2.3 Ciclo de vida del crédito

La labor de las instituciones financieras sobre los créditos no termina en el otorgamiento o rechazo del mismo, también se lleva a cabo su análisis sobre su ciclo de vida. Cabe recalcar que la principal razón por la que se estudia este ciclo es para tener un buen manejo sobre el cliente y así saber por medio de los pagos que realiza si se espera que termine de pagar todo el crédito o en caso contrario que el usuario incumpla con sus obligaciones establecidas, no obstante también ayuda para dar promociones con respecto a los créditos a los usuarios que han tenido

un buen comportamiento sobre sus pagos y de esta manera generar mejores ingresos a las instituciones.

Figura 1. *Ciclo de vida del crédito*.

Admisión

La admisión es la primera fase del ciclo de vida del crédito y se caracteriza por ofrecer el crédito, revisar la documentación y en aceptar o rechazar la solicitud del cliente según sus características. La aceptación o rechazo de la solicitud se realiza por medio de los modelos de calificación crediticia la cual habrá sido tratada con información de la población objetivo y el nivel de riesgo que la institución quiera aceptar.

Seguimiento

El seguimiento es la segunda fase del ciclo en donde se realiza el servicio al cliente y se evalúa la puntualidad de los pagos, en este caso si se realizó en tiempo y forma el pago o existió morosidad. Esta fase es muy importante ya que se va evaluando al usuario por medio de cada pago ya que si ha realizado todos los pagos a tiempo existe mayor posibilidad de que termine de pagar todo mientras que; por otro lado, si el cliente no ha pagado más de una anualidad se vuelve más complicado que éste lo termine de pagar y por lo tanto la probabilidad de que se pague por completo es menor.

El seguimiento es una función básica que ayuda a garantizar un adecuado tratamiento al riesgo en función de su estado, un claro ejemplo está en donde el cliente no ha pagado y debido a esto

se realizan varios métodos de rastreo para impulsarlo a que cumpla con su responsabilidad y en el peor de los casos en donde se da por perdido el pago.

Cobranza o liquidación

La última fase del ciclo de vida de un crédito se divide en dos, ya que no siempre el cliente cumplirá con sus responsabilidades, se le denomina liquidación cuando al final del tiempo estipulado se paga todo el crédito y cobranza cuando el usuario no concluyó con el pago.

Cuando existe la liquidación, el banco generalmente le sigue ofreciendo al cliente créditos, ya que ha tenido un buen comportamiento con respecto de sus pagos, del lado contrario tenemos que cuando existe la cobranza el banco toma estrategias de recuperación para no tener pérdidas considerables sobre la inversión realizada, las cuales van encaminadas a recuperar la deuda en el menor tiempo posible valorando la alternativa económica más beneficiosa.

1.3 El riesgo

En la sección anterior nos percatamos que el factor más importante en el rechazo o aprobación de un crédito es el riesgo, acerca del riesgo el Banco de México (BANXICO, 2015) afirma que;

> Está relacionado con la posibilidad de que ocurra un evento que se traduce en pérdidas para los participantes en los mercados financieros, como pueden ser inversionistas, deudores o entidades financieras. El riesgo es producto de la incertidumbre que existe sobre el valor de los activos financieros, ante movimientos adversos de los factores que determinan su precio; a mayor incertidumbre mayor riesgo.

Como se mencionó en la definición de riesgo, es un daño, siniestro o pérdida que ninguna persona o institución tiene intención de que le suceda, que puede que se materialice o no. El riesgo siempre ha existido y existirá, dado que es inevitable en todos los procesos de toma de decisión.

1.3.1 Clasificación del riesgo

Los riesgos a los que una institución financiera está expuesta son de una diversidad grande y por lo tanto se pueden clasificar de diferentes maneras, en este trabajo se clasificaron de acuerdo a los estándares regulatorios del sistema financiero mexicano establecidos por la Comisión Nacional Bancaria y de Valores [CNBV]:

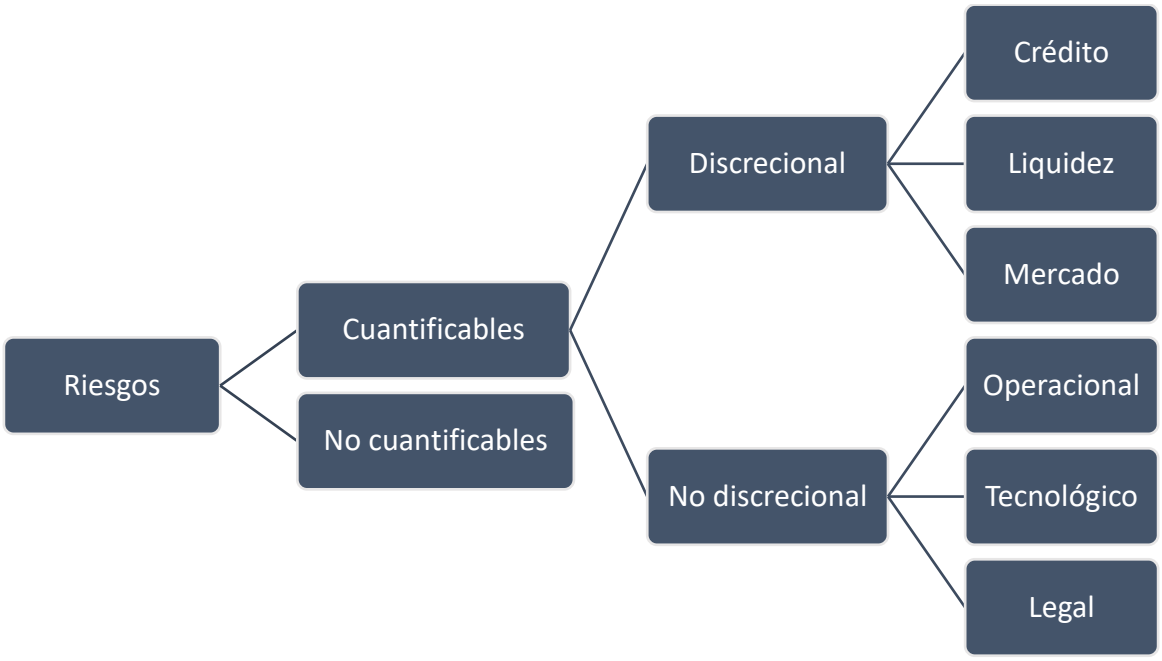


Figura 2 . *Esquema de la clasificación de riesgos según la CNBV*.

Riesgos cuantificables: “son aquéllos para los cuales es posible conformar bases estadísticas que permitan medir sus pérdidas potenciales” (CNBV).

En este tipo de riesgo se tiene más clara la intensidad que puede llegar a ocasionar si se materializa el riesgo. Gracias a los métodos estadísticos de hoy en día, el riesgo cuantificable es en donde recaen la mayoría de los tipos de riesgos, este se divide en dos, uno es el riesgo discrecional y por su contraparte se encuentra el no discrecional.

“**Riesgos discrecionales**, que son aquellos resultantes de la toma de una posición de riesgo” (CNBV, p. 110).

Este riesgo se divide en 3, riesgo de crédito, riesgo de liquidez y riesgo de mercado, estos riesgos son los principales en los que las instituciones financieras se enfocan y a continuación veremos las razones por las cuales lo hacen.

> “**Riesgo de crédito o crediticio**, que se define como la pérdida potencial por la falta de pago de un acreditado o contraparte en las operaciones que efectúan las Instituciones, incluyendo las garantías reales o personales que les otorguen, así como cualquier otro mecanismo de mitigación utilizado por las Instituciones” (CNBV).

El riesgo de crédito es inminente para las instituciones financieras y es la principal razón por el cual se realiza este trabajo, ya que se deben de tomar acciones para poner disminuirlo y los modelos de calificación crediticia son perfectos para ello.

> “**Riesgo de liquidez**, que se define como la pérdida potencial por la imposibilidad o dificultad de renovar pasivos o de contratar otros en condiciones normales para la Institución, por la venta anticipada o forzosa de activos a descuentos inusuales para hacer frente a sus obligaciones, o bien, por el hecho de que una posición no pueda ser oportunamente enajenada, adquirida o cubierta mediante el establecimiento de una posición contraria equivalente” (CNBV).

Para poder entender bien el riesgo de liquidez, primero debemos entender el concepto de liquidez lo cual, lo expresó de buena manera Cuartas (2013): “Se considera la liquidez como la posibilidad inmediata y directa de conversión del dinero efectivo en cualquier tipo de bienes materiales de valor. La liquidez permite tener poder de compra o de pago inmediato”.

Este tipo de riesgo es interesante dado que la institución sufre de este riesgo por ausencia de dinero disponible, pero por lo otro lado si existe demasiado dinero disponible, la institución sufre pérdidas, ya que tiene estancado dinero de más, el cual podría estar generando una ganancia.

De la misma manera Lara (2008) menciona que “Este riesgo se presenta en situaciones de crisis, cuando en los mercados hay únicamente vendedores.”

Riesgo de mercado, que se define como la pérdida potencial por cambios en los Factores de Riesgo que inciden sobre la valuación o sobre los resultados esperados de las operaciones activas, pasivas o causantes de pasivo contingente, tales como tasas de interés, tipos de cambio e índices de precios, entre otros (CNBV).

Este riesgo hace referencia a los posibles cambios en el mercado, que ocasionen pérdidas de valor de un activo. Un ejemplo de riesgo de mercado que hace referencia al tipo de cambio es cuando una empresa compra su materia prima fuera del país y existe una variación no esperada en el tipo de cambio.

Riesgos no discrecionales, son aquéllos resultantes de la operación del negocio, pero que no son producto de la toma de una posición de riesgo, tales como el **riesgo operacional**, que se define como la pérdida potencial por fallas o deficiencias en los controles internos, por errores en el procesamiento y almacenamiento de las operaciones o en la transmisión de información, así como por resoluciones administrativas y judiciales adversas, fraudes o robos, y comprende, entre otros, al riesgo tecnológico y al riesgo legal (CNBV).

Los riesgos no discrecionales se dividen en tecnológico y legal.

El riesgo tecnológico se define como la pérdida potencial por daños, interrupción, alteración o fallas derivadas del uso o dependencia en el hardware, software, sistemas, aplicaciones, redes y cualquier otro canal de distribución de información en la prestación de servicios bancarios con los clientes de la Institución (CNBV).

Esta es una de las razones por las que se eligió la clasificación de riesgos de la CNBV, dado que este riesgo ha surgido en los últimos años y hoy en día la tecnología es vital para las personas y más cuando se trata de la banca, ya que por medio de las aplicaciones del banco se realizan diferentes acciones como pagar servicios, mandar dinero y analizar saldos de cuenta, por lo tanto, también las instituciones deben de analizar y disminuir este tipo de riesgos.

El riesgo legal se define como la pérdida potencial por el incumplimiento de las disposiciones legales y administrativas aplicables, la emisión de resoluciones administrativas y judiciales desfavorables y la aplicación de sanciones, en relación con las operaciones que las Instituciones llevan a cabo (CNBV).

El otro tipo de riesgo, contrario al cuantificable es “**Riesgos no cuantificables**, que son aquéllos derivados de eventos imprevistos para los cuales no se puede conformar una base estadística que permita medir las pérdidas potenciales” (CNBV).

El peligro de este riesgo es que no se sabe con certeza el grado de gravedad que podría llegar a ocasionar, por lo tanto, se sabe que se mide en términos cualitativos.

1.3.2 Riesgo de crédito

Dada la importancia que tiene el riesgo de crédito en este trabajo se realizó un mayor énfasis y se tiene que se divide en dos; el riesgo de contraparte y el riesgo de crédito.

Riesgo de contraparte

> Existe cuando se da la posibilidad de que una de las partes de un contrato financiero sea incapaz de cumplir con las obligaciones financieras contraídas, haciendo que la otra parte del contrato incurra en una pérdida (Banxico).

El riesgo de contraparte hace referencia a lo ya antes mencionado, cuando una de las partes del contrato no cumple con su obligación de pagar el crédito.

Riesgo de crédito

> Es el caso particular cuando el contrato es uno de crédito, y el deudor no puede pagar su deuda. Recientemente, además del caso de incumplimiento, se han incorporado eventos que afectan el valor de un crédito, sin que necesariamente signifique incumplimiento del deudor. Esto ocurre típicamente por cambios en la calidad de un crédito, cuando una calificadora lo degrada. Cuando esto ocurre, significa que la calificadora considera que ha aumentado la probabilidad de incumplimiento del emisor de la deuda, y por lo tanto el crédito vale menos ya que se descuenta a una tasa mayor (Banxico).

Con la definición anterior nos podemos dar cuenta que es más completa que las anteriormente dadas y así de esta manera analizar también que la tasa a la que está referenciado el crédito puede afectar en su calidad tanto para el que lo solicita como el que lo otorga.

En este punto del trabajo ya entendemos la definición de riesgo de una manera totalmente teórica y para introducir la parte práctica del tema existen un conjunto de elementos que ayudan a medir el riesgo de crédito y según Banxico son:

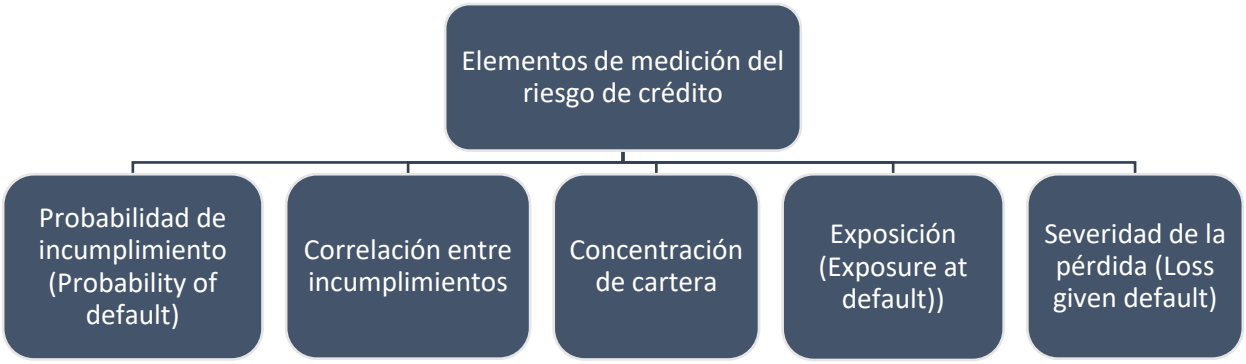


Figura 3. *Elementos de medición del riesgo de crédito.*

Probabilidad de incumplimiento (PD)

> Es la medida de qué tan probable es que un acreditado deje de cumplir con sus obligaciones contractuales … Por tipo de crédito, normalmente se estima a partir de la tasa de incumplimiento observada en cada tipo de crédito, que es la proporción de deudores o créditos que dejan de pagar en un periodo de tiempo dado, respecto de los que estaban vigentes en el periodo anterior (Banxico, 2005).

Por definición, nosotros tenemos que es una probabilidad, por lo tanto, se tiene que cero es su valor mínimo haciendo referencia a que es imposible que incumpla con sus obligaciones y uno, cuando es seguro que incumpla.

Correlación entre incumplimientos

“La correlación “a pares” mide la dependencia o grado de asociación entre el comportamiento crediticio de dos deudores“ (Banxico, 2005).

Hace referencia a la relación que existe entre dos clientes, el valor de la correlación va de -1 a 1, tomando en cuenta que en 1 es cuando existe la máxima relación positiva y en -1 la máxima correlación negativa, un ejemplo en donde existe correlación positiva entre dos clientes es cuando un cliente deja de pagar y el otro cliente también cae en incumplimiento del pago y por el otro lado, cuando existe la máxima correlación negativa, si uno paga el otro se espera que no pague.

Concentración de cartera.

> Concentración significa que hay mucho crédito en pocas manos, lo cual puede ser riesgoso. La concentración se puede dar en muchos sentidos y es más peligrosa cuando se da en segmentos riesgosos de la cartera … Un indicador muy conocido para medir la concentración es el índice de Herfindahl Hirshmann (IHH) que toma valores entre el recíproco del número de deudores o créditos (N) de una cartera, y uno. Así, una cartera totalmente diversificada en donde todos los deudores deben exactamente lo mismo daría un valor del índice de 1/N, mientras que, si el índice vale uno, necesariamente se tiene que el crédito se encuentra totalmente concentrado en un solo crédito o deudor (Banxico, 2005).

Este elemento es muy importante en la selección de los clientes, dado que, si se llega tener una cartera crediticia no diversificada puede ser muy riesgoso, un ejemplo claro puede ser cuando una institución otorga pocos créditos con sumas aseguradas muy altas, si pocos de los clientes dejan de pagar sus deudas, se vería afectada de una manera notable la institución, ya que la mayor parte del dinero no será pagada, por otro lado cuando existe la diversificación, la institución podría afrontar de una mejor manera en que varios deudores dejen de pagar, porque habrán otros más que si paguen y se espera que la proporción de la deuda que no se pagó no sea significativa con respecto de la que sí se pagó.

Exposición (EAD)

“Es lo que debe el deudor en un momento dado en caso de incumplimiento” (Banxico, 2005).

De cierta manera nos podemos dar cuenta que esta definición es intuitiva, dado que, al inicio de la deuda existe mayor exposición por el hecho de que no se ha amortizado la mayor parte de lo que se debe, por otro lado, cuando el cliente se encuentra realizando los últimos pagos existe una exposición mucho menor porque ya se realizaron la mayoría de estos.

Severidad de la pérdida (LGD)

Para poder entender la definición de severidad de la pérdida, de antemano explicaremos otra definición como apoyo, la cual es la tasa de recuperación del crédito y se entiende como la proporción del crédito que podrá ser recuperada una vez que la contraparte ha caído en el incumplimiento.

> Esto es lo que pierde el acreedor en caso de incumplimiento del deudor y se mide como una proporción de la exposición … la severidad representa el costo neto del incumplimiento de un deudor; es decir, la parte no recuperada al incumplir el acreditado una vez tomados en cuenta todos los costos implicados en dicha recuperación (Banxico, 2005).

Por medio de este término nos podemos dar cuenta el grado de severidad que tiene el riesgo si se llegase a materializar, tomando en cuenta que la severidad de la pérdida es la proporción de la exposición del crédito tenemos que LGD será mayor a cero y menor o igual a uno, se comprende que LGD es menor o igual a uno dado que el cliente puede no pagar el crédito desde el primer periodo y por el otro lado es mayor a cero dado que por definición tenemos que el cliente cae en incumplimiento durante el periodo.

Cuando se trata de la distribución de pérdidas y ganancias, existe un conjunto de cálculos básicos que son indispensables para aproximar el nivel riesgo, los cuales son: Value at Risk, pérdida esperada y pérdida no esperada.

Value at Risk (VaR)

> Al igual que en riesgo de mercado, el valor en riesgo de una cartera de crédito es el cuantil de la distribución de pérdidas y ganancias asociada a la cartera de crédito, para el periodo de

tiempo y el nivel de confianza escogidos. Normalmente se descompone en lo que se conoce como la pérdida esperada y la no-esperada (Banxico, 2005).

Se entiende al VaR como la máxima perdida posible en un tiempo especificado y se calcula por medio de la serie de tiempo de la variable a evaluar ajustando dos valores, el periodo de tiempo y el nivel de confianza, el ultimo mencionado sirve para medir la confianza del modelo, en la mayoría de los casos se utiliza el 5%, mientras más pequeño sea el valor, más confiable será.

Pérdida esperada (PE)

Es la media de la distribución de pérdidas y ganancias, es decir, indica cuánto se puede perder en promedio y normalmente está asociada a la política de reservas preventivas que la institución debe tener contra riesgos crediticios (Banxico, 2005).

Generalmente la pérdida esperada tiene un horizonte de tiempo de 12 meses y se calcula de la siguiente manera:

$$PE = PD \cdot LGD \cdot EAD$$

Donde

PE: Pérdida esperada

PD: Probabilidad de incumplimiento

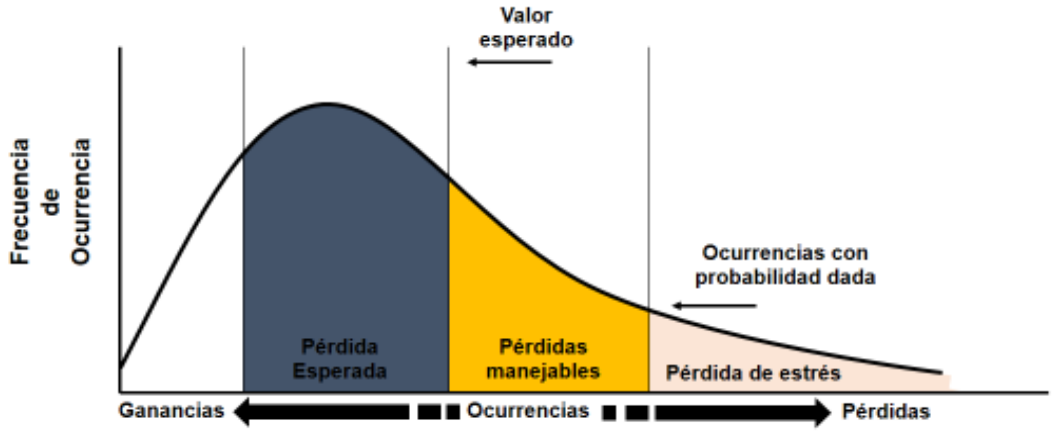
LGD: Severidad de la pérdida

EAD: Exposición al momento del incumplimiento

Pérdida no esperada

Es la pérdida por encima de la esperada, medida como el $VaR - PE$, en que puede incurrir el acreedor, por incumplimiento de sus deudores … estas pérdidas determinan el capital económico requerido por el acreedor para hacer frente a pérdidas no anticipadas (Banxico, 2005).

La pérdida no esperada se puede comprender como el promedio de las pérdidas una vez que se haya superado la pérdida esperada.



Nivel de exposición	Costo Esperado	Pérdida máxima aceptable	Eventos extremos
Metodología de la exposición	Pérdida esperada	VaR	Análisis de estrés
Acciones de control	Estados de resultados	Límites en VaR	Límites a pérdidas por estrés

Figura 4. *Medidas de riesgo*.

La Figura 4 analiza de izquierda a derecha, donde en el eje x se encuentran las ocurrencias de las pérdidas y en el eje y las frecuencias, por ello, mientras más hacia la derecha se encuentren los valores existirán mayores pérdidas. El análisis de la definiciones anteriormente dadas se realizan de la misma manera, de izquierda a derecha tomando en cuenta que el primer aspecto mencionado es la pérdida esperada, donde generalmente podemos encontrar la mayor parte de las pérdidas, es por ello que en esta sección se encuentra la parte más alta de la gráfica (donde mayor frecuencia existe), cuando las pérdidas sobrepasan este nivel de ocurrencia se le denominan pérdidas no esperadas la cuales se encuentran en la parte de la derecha de la gráfica y entre más hacia la derecha se encuentren, son más catastróficas.

1.4 Proceso de administración de riesgos

Todas las acciones que realiza una empresa tomando en cuenta el riesgo que conllevan sus operaciones se les pueden denominar administración de riesgos y Lara (2008) argumentó sobre ello:

> La medición efectiva y cuantitativa del riesgo se asocia con la probabilidad de una pérdida en el futuro. Los seres humanos deben conocer y responder de manera intuitiva o cuantitativa a las probabilidades que confrontan en cada decisión. La esencia de la administración de riesgos consiste en medir esas probabilidades en contextos de incertidumbre.

El cálculo de probabilidades sobre eventos futuros sirve como un gran sustento en la toma de decisiones de las personas, dado que, si queremos que ocurra un evento y dicho evento tiene una probabilidad baja de que ocurra, entonces se le asigna un riesgo alto, lo que ocasiona que busquemos medidas para poder reducir este riesgo o en todo caso rechazar la oferta y así poder evitar pérdidas potencialmente significativas. En este tipo de acciones es en lo que consiste la administración de riesgos.

¿Qué es la administración de riesgos? Según la CNBV es: “El conjunto de objetivos, políticas, procedimientos y acciones que se llevan a cabo para identificar, medir, vigilar, limitar, controlar, informar y revelar los distintos riesgos a que se encuentran expuestas las Instituciones.” En resumen, es un proceso para poder analizar y afrontar los riesgos que sufre una institución.

Los objetivos más importantes de la administración de riesgo según de Lara (2008) son:

- Asegurarse de que una institución o inversionista no sufra pérdidas económicas inaceptables (no tolerables).
- Mejorar el desempeño financiero de dicho agente económico, tomando en cuenta el rendimiento ajustado por riesgo.

De acuerdo con lo mencionado, este procedimiento es de suma importancia para toda empresa, como se comentó en el primer objetivo, ayuda a que la empresa no tenga pérdidas que afecten notablemente el funcionamiento de una institución, esto ayuda para tener un mejor rendimiento, sin contar que de esta manera la empresa no caiga con mayor facilidad en crisis.

De la misma manera Basilea (1999) comentó esto acerca de la administración de riesgos: “El objetivo de la administración del riesgo de crédito es maximizar la tasa de rendimiento ajustada por el riesgo del banco, manteniendo la exposición al riesgo de crédito dentro de límites aceptables“ Esta idea complementó de buena manera al segundo objetivo que argumentó de Lara, la administración de riesgos mejora el desempeño financiero al máximo, tomando en cuenta el riesgo que conlleva esto, lo que provoca tener el mejor desempeño posible siendo económicamente estables.

Como se aprecia en los objetivos anteriores la administración de riesgos es una práctica que deben de realizar todas las instituciones e inversionistas, dado que es vital para tener un buen funcionamiento en el manejo de las inversiones.

> El proceso de administración de riesgos implica, en primer lugar, la identificación de riesgos, en segundo su cuantificación y control mediante el establecimiento de límites de tolerancia al riesgo y finalmente: la modificación o nulificación de dichos riesgos a través de disminuir la exposición al riesgo o de instrumentar una cobertura (de Lara, 2008).

A continuación, en la Figura 5 se presenta de manera jerárquica el orden en el que se realiza un adecuado proceso de administración de riesgos:

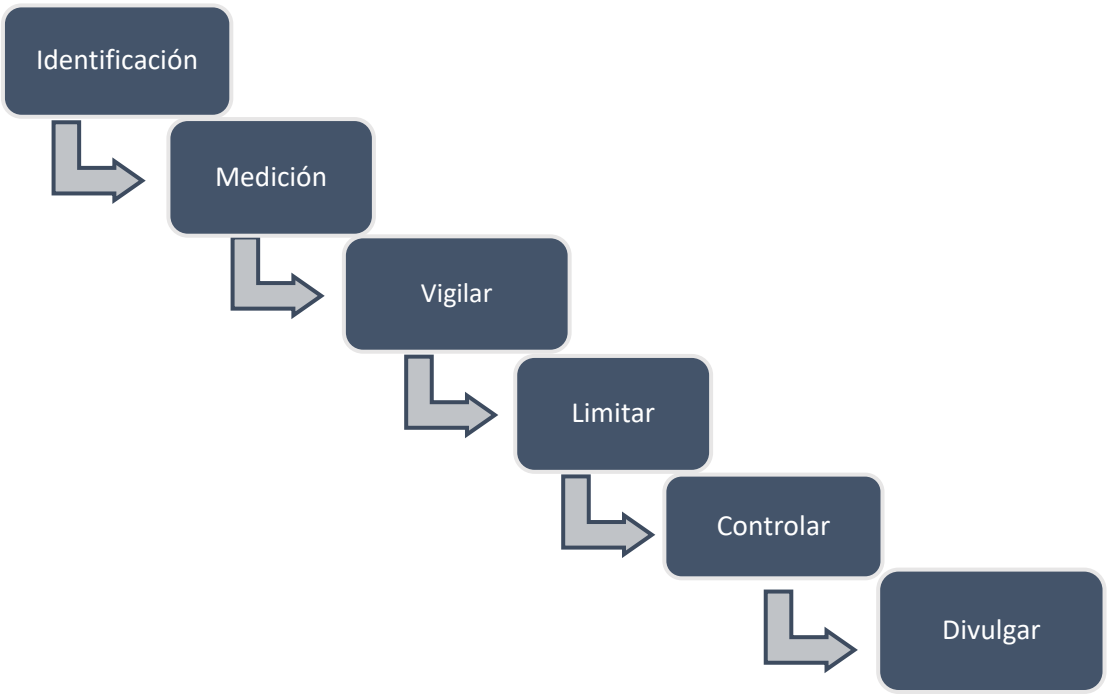


Figura 5. *Proceso de administración de riesgos*.

Identificación, identificar los riesgos a los que está expuesta la institución o el inversionista en sus operaciones, esta fase abarca todos los riesgos desde un incendio en el lugar de trabajo hasta las pérdidas ocasionadas por malas inversiones.

Medición, medir los riesgos mediante métodos estadísticos para saber el impacto que tendría en la institución si se materializara el riesgo.

Vigilar, las tendencias del riesgo y su comportamiento.

Limitar, la exposición del riesgo conforme a los objetivos establecidos.

Controlar, eficazmente las posiciones activas y pasivas.

Divulgar, a los órganos tomadores de decisiones los riesgos asumidos, la rentabilidad ajustada por riesgo y las medidas propuestas para reducir el riesgo.

La divulgación de información es una de las fases más importantes del proceso por varias razones, Vandemaele, Vergauwen, & Michiels, (2009) explican cuáles son:

> Mitiga la asimetría de información entre la administración y los accionistas externos y puede tener efectos positivos en la confianza que las partes interesadas tienen en la administración de la empresa. Puede disminuir el riesgo percibido de la empresa porque una estrategia de divulgación abierta supuestamente da como resultado una mejor evaluación del desempeño futuro de la empresa. Esto, a su vez, puede conducir a una disminución en el costo de capital de la empresa.

Todas las fases que conlleva la administración de riesgos son importantes y no hay que perder de vista ninguna de éstas, un ejemplo puede ser que no se haga un buen análisis acerca del problema y al final se divulgue información que no sea correcta y por lo tanto perjudique a la institución u otro ejemplo que no haya una buena divulgación de la información, lo que genera que muy poca gente conozca sobre la solución encontrada.

Como se había mencionado al principio del capítulo, el banco es una institución que sufre algunos riesgos y en particular el de crédito, a continuación, la CNBV (2017) comentó algunas medidas que los bancos realizan como tarea de la administración de riesgos:

> Reservas para hacer frente a los riesgos implícitos en la operación. Administrar adecuadamente el riesgo implícito en sus operaciones. Gestión de activos y pasivos (diversificación, control de liquidez, etc.) Calificación de cartera de crédito. Políticas sólidas de otorgamiento de crédito. Vigilancia y seguimiento de la cartera vencida y recuperación de cartera.

1.5 Desastres financieros ocasionados por la ausencia de administración de riesgos

Una manera de entender la gran importancia que tiene la administración de riesgos en las instituciones es por medio de la historia, a continuación, se presentan en la Figura 6 los desastres financieros más conocidos por la ausencia de la administración de riesgos:

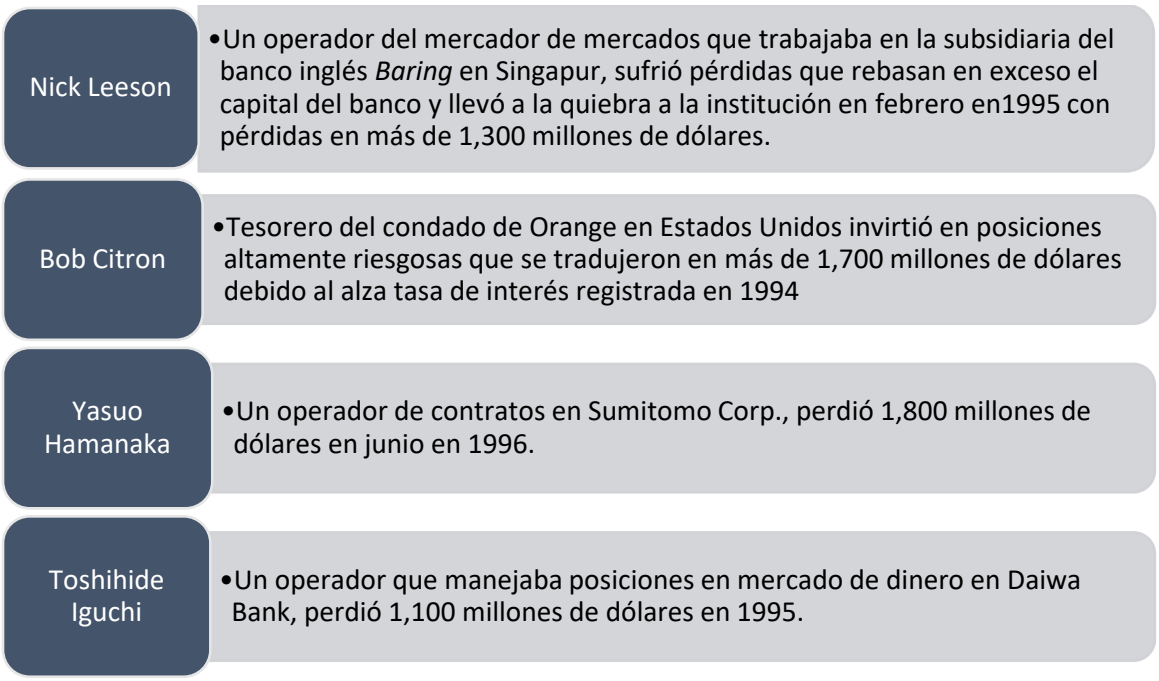


Figura 6. *Desastres financieros.*

Los sucesos anteriormente presentados son un claro ejemplo de la relevancia que tiene el proceso de administración de riesgos en una institución, ya que al no tener en cuenta el riesgo que conlleva cada acción es imposible saber si puede llegar a ser peligroso, de la misma manera se concluye que el costo que se puede llegar a pagar por omitir este tipo de acciones puede llegar a ser demasiado alto, al punto de llevar a la ruina la empresa.

Capítulo 2.- Modelos de clasificación

En este capítulo se presenta la información para entender la elaboración de un clasificador empezando por el conjunto de datos a evaluar, porque se segmenta en dos partes, la primera que sirve para entrenar al modelo y la segunda que se utiliza para su evaluación, también otro aspecto que se analiza es en los tipos de clasificadores ya que existen dos, uno donde las únicas salidas que tiene el modelo son las predefinidas por medio de los datos de entrenamiento y el otro en donde el mismo clasificador segmenta a la población de acuerdo a su criterio, por último se presenta la parte teórica de los clasificadores así como de las pruebas utilizadas para evaluarlos.

2.1 Introducción a los modelos de clasificación

A lo largo de este capítulo se presentan los clasificadores más relevantes que existen para segmentar a las personas que tienen más probabilidad de pagar un préstamo con las que no, pero ¿sabemos a qué hace referencia el término clasificación?

Clasificación es el acto o proceso de asignar un objeto en el lugar que corresponda dentro de un conjunto de categorías establecido. Los atributos básicos de cada categoría son conocidos, aunque haya algunas incertidumbres a la hora de asignar alguna observación dada (Lara Albín, 2014).

En resumen, se define la clasificación como el proceso de identificar observaciones nuevas y asignarles una categoría por medio de sus características. El autor en el párrafo anterior hizo énfasis en que existe incertidumbre en realizar esta actividad, dado que, a pesar de tener información sobre los atributos de cada observación, existe probabilidad de que la opción elegida no sea la correcta porque no se sabe con certeza a qué categoría pertenece.

Antes de presentar los modelos, es necesario explicar las partes en las que se puede segmentar al conjunto de datos para la correcta elaboración del clasificador:

Datos de entrenamiento

Los datos de entrenamiento o de ejemplo, como su nombre lo indica, son un conjunto de datos que utiliza el modelo para comprender la relación que existe entre los atributos de las observaciones y la categoría a la que pertenecen y en base a esto clasificar las observaciones futuras, cabe recalcar que se espera que las nuevas observaciones a clasificar tengan un comportamiento similar a los datos de entrenamiento.

Datos de prueba

Los datos de prueba se emplean después de entrenar al clasificador y sirven para evaluar el nivel de predicción que tiene dicho modelo, dado que se tiene el conocimiento de los atributos de cada observación y de la categoría a la que pertenecen, podemos hacer la comparación en si la respuesta del modelo concuerda con las categorías que pertenecen.

Ahora que sabemos que el conjunto de datos a estudiar se puede dividir en dos, se sabe que, generalmente el 80% de las observaciones son de entrenamiento y el 20% restante se utiliza como prueba.

Los métodos de clasificación son técnicas matemáticas enfocadas al análisis de los atributos de cada observación para su categorización. Para tener un mayor entendimiento sobre este tipo de métodos a continuación se presenta la clasificación de los modelos estadísticos que se utilizaron a lo largo de este documento.

Modelos de clasificación

A continuación, en la Figura 7 se presenta la forma en que se segmentaron a los modelos de clasificación.

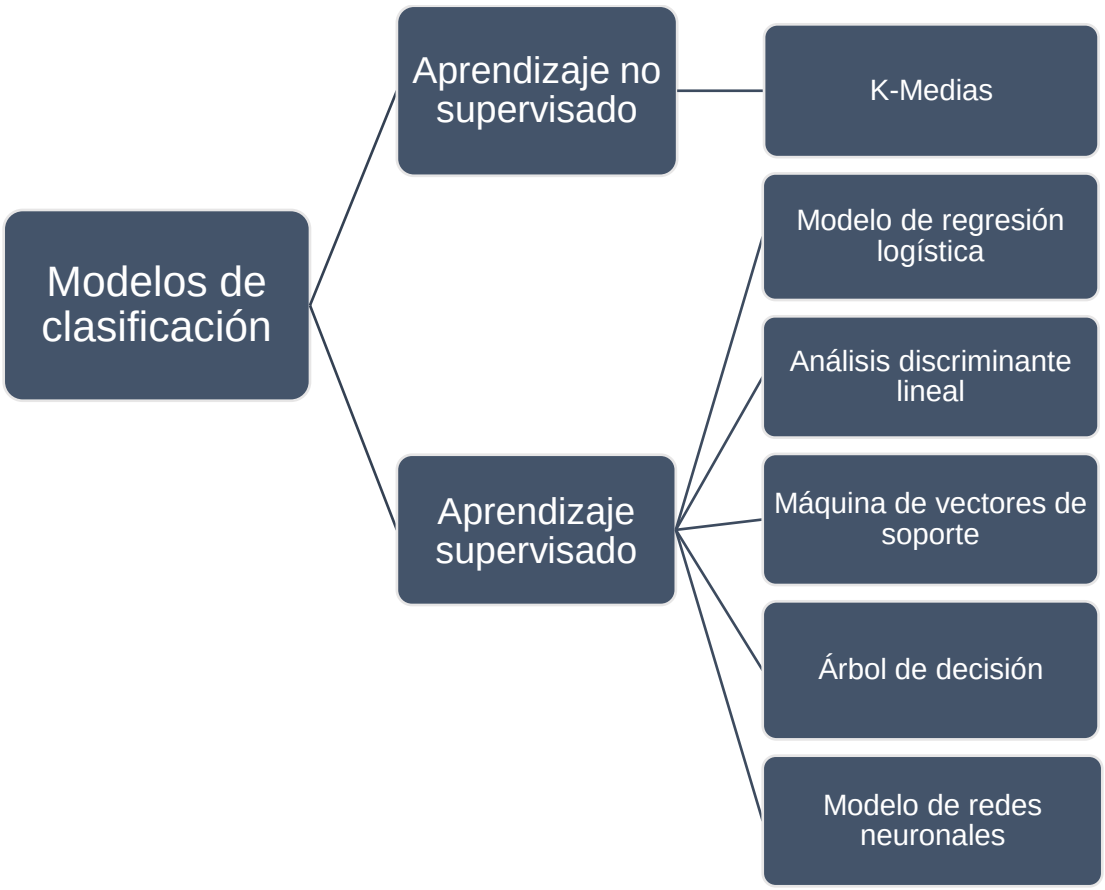


Figura 7. *Modelos de clasificación.*

Aprendizaje supervisado

En este tipo de aprendizaje se utilizan los datos de entrenamiento, ya que por medio de estos el analista de datos le dice al modelo cuales son las posibles categorías por clasificar y cuales deben de ser sus atributos, algunos autores han comentado sobre el aprendizaje supervisado que:

> Es un tipo de aprendizaje automático que se basa en que una máquina pueda aprender una manera de convertir determinada entrada en una salida basada en un conjunto de datos de ejemplo (Páez, 2019).

En resumen, tenemos que es un tipo de aprendizaje, por el cual el modelo trata de aprender a categorizar las nuevas observaciones por medio de los datos de ejemplo y tomando en cuenta que las categorías en las que serán clasificadas las observaciones están predefinidas.

Algunos modelos del aprendizaje supervisado son:

- Modelo de regresión logística.
- Análisis discriminante lineal.
- Máquina de vectores de soporte.
- Árbol de decisión.
- Clasificación de Naive Bayes.
- Regresión por mínimos cuadrados.
- Validación cruzada.

Aprendizaje no supervisado

Algunos autores han comentado que en el aprendizaje no supervisado se tiene que: “la máquina simplemente recibe las entradas $x_1, x_2, x_3, \ldots$, pero sin ningún conjunto de salidas deseadas” (Páez, 2019).

A diferencia del aprendizaje supervisado, el aprendizaje no supervisado no tiene ningún conocimiento previo de lo que se quiere aprender, en pocas palabras no se define a las categorías deseadas, si no que el mismo modelo propone las categorías en las cuales serán clasificadas las nuevas observaciones.

De primera instancia, tal vez sea extraño en que una persona quiera hacer una clasificación de un conjunto de datos, sin saber de antemano las categorías en las que serán clasificadas, pero James et al. (2013), dio un ejemplo sobre un conjunto de datos de *n* observaciones que corresponde a pacientes con cáncer de mama y *p* características correspondientes de las mediciones recolectadas:

> Es posible que tengamos una razón para creer que existe cierta heterogeneidad entre las n muestras de tejido; por ejemplo, tal vez haya algunos subtipos desconocidos diferentes de cáncer de mama. La agrupación podría utilizarse para encontrar estos subgrupos (James et al., 2013).

En este ejemplo no se introducen las categorías en las que se quiere segmentar al conjunto de datos, el modelo parte de las mismas características de las observaciones, dado que gracias a estas se van a ir segmentando con las que más coincidan entre sí. Si se hubiera realizado este ejercicio por medio de un aprendizaje supervisado podría existir un error ya que es posible que se omita alguna clase de cáncer.

Algunos ejemplos de aprendizaje no supervisado son:

- Modelo de k- medias.
- Agrupamiento jerárquico.
- Modelo de agrupamiento gaussiano.
- Mapas auto- organizados.

Existe una gran variedad de modelos de clasificación, ya que se ocupan en las diferentes ramas de la ciencia, pero en la siguiente sección se presentan de una manera resumida los modelos más relevantes para la clasificación de créditos.

2.2 Modelos de aprendizaje supervisado

2.2.1 Modelo de regresión logística

Este modelo estadístico es uno de lo más usados en la actualidad por su fácil entendimiento y predicciones aceptables, Días Monrroy & Morales Rivera (2012) comentaron sobre ello: “Es un caso especial del modelo de regresión, donde el criterio de respuesta es de tipo categórico o discreto. El interés se dirige a investigar los efectos de un conjunto de predictores sobre la respuesta, las variables predictoras pueden ser de tipo cuantitativo, categórico o de ambas”.

Tenemos que la variable dependiente es binaria, es decir solo puede tomar dos valores y por comodidad en este trabajo dicha variable solo puede tomar el valor cero o uno.

Sea Y la variable dependiente y $X = (X_1, \dots, X_k)$, los atributos de cada observación, entonces tenemos que la probabilidad de que la variable dependiente sea igual a uno, dado los atributos de las observaciones es:

$$P(Y = 1|X) = P(Y = 1|X_1, \dots, X_k) = f(\beta_0 + x_1\beta_1 \ + \cdots + \ x_k\beta_k \ ) \quad \beta_1, \dots, \beta_k \, \varepsilon \, R$$

Donde $\beta_1, \dots, \beta_k$ es el peso que tiene cada atributo respecto a la variable dependiente, dichos valores son desconocidos y por lo tanto se deben de estimar.

Para este clasificador sólo se usa la función logística para obtener la probabilidad de que dicha muestra pertenezca a cierto segmento:

$$f(\beta_0 + x_1\beta_1 \ + \cdots + \ x_k\beta_k \ ) = \frac{1}{1 + e^{-(\beta_0 + x_1\beta_1 \ + \cdots + \ x_k\beta_k \ )}}$$

La función Logística es estrictamente creciente, y cuando X tiende a $-\infty$, es decir cuando toma números grandes negativos, entonces f(x) → 0, por otro lado, si x tiende a ∞ entonces f(X) → 1. A continuación, en la Figura 8 se presenta la gráfica de la función logística y se aprecia como ésta toma forma de una “S”:

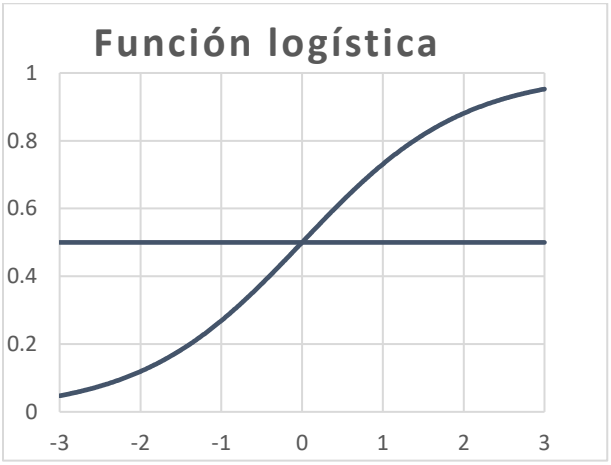


Figura 8. *Gráfica de la función logística*

Queremos saber el resultado de la variable dependiente dado que, tenemos las variables explicativas, entonces debemos de encontrar la función inversa, para ello primero se tiene que demostrar que dicha función existe y Armando et al (2008) comentó que la definición de función inversa es: "Sea $f: A \rightarrow B$ una función biyectiva. La función $g: B \rightarrow A$ definida mediante la regla $g(b) = a$, donde a es tal que $f(a) = b$, se le denomina función inversa de f".

Entonces se tiene que demostrar que la función logística es biyectiva, es decir que es inyectiva y suprayectiva:

P.D. que la función logística es inyectiva:

$$f(z_1) = f(z_2) \rightarrow \frac{1}{1+e^{-(z_1)}} = \frac{1}{1+e^{-(z_2)}} \rightarrow 1 + e^{-(z_1)} = 1 + e^{-(z_2)} \rightarrow e^{-(z_1)} = e^{-(z_2)} \rightarrow z_1 = z_2$$

P.D que la función logística es suprayectiva:

$$f(z) = f\left(\ln\left(\frac{f(z)}{1-f(z)}\right)\right) = \frac{1}{1+e^{-\ln\left(\frac{f(z)}{1-f(z)}\right)}} = \frac{1}{1+\left(\frac{1-f(z)}{f(z)}\right)} = f(z)$$

Por lo tanto, la función inversa es:

$$f(z) = \frac{1}{1+e^{-z}} <=>$$

$$\frac{1}{f(z)} = 1 + e^{-z} <=>$$

$$\frac{1}{f(z)} - 1 = e^{-z} <=>$$

$$\frac{1-f(z)}{f(z)} = e^{-z} <=>$$

$$\frac{f(z)}{1-f(z)} = e^{z} <=>$$

$$\ln\left(\frac{f(z)}{1-f(z)}\right) = z = (\beta_0 + x_1\beta_1 + \cdots + x_k\beta_k)$$

La última ecuación es conocida como log- odds o logit.

Como sabemos, tenemos que esta función nos arrojará valores entre cero y uno, pero ¿Cómo convertimos estos valores a un resultado binario donde solo arroje un cero o un uno? esto se hace por medio de un umbral, sea y^* dicho umbral, entonces si $P(Y = 1|X) > y^*$ entonces el clasificador arrojará un 1 de lo contrario será 0.

Estimación de los parámetros

Tomando en cuenta las ecuaciones anteriormente presentadas, nos enfrentamos a un problema, que es encontrar el valor estimado para cada coeficiente ($\beta_0, \beta_1 , \dots, \beta_k$) que hace referencia al peso que tiene cada variable explicativa con respecto de la variable dependiente y estos valores deben de ser estimados por medio del conjunto de datos a utilizar.

Existen varios métodos para estimar los parámetros, pero según James, Witten, Hastie, & Tibshirani (2013) recomiendan que: “el método general de máxima verosimilitud, ya que tiene mejores propiedades estadísticas”

Dado que la variable **Y** solo toma dos valores y se supone que el resultado de la regresión logística es la probabilidad de que la variable dependiente sea uno, podemos considerar a la función como una variable aleatoria Bernoulli: $Y \sim Ber(z)\ donde\ z = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k$

$$g(y) = P(Y = y|\ X)\ = f(z)^y \cdot (1 - f(z))^{1-y}\ , \qquad y = 0{,}1$$

Se considera que las *k* observaciones son independientes:

$$L(\theta) = \prod_{i=1}^{k} g_i(y) = \prod_{i=1}^{k} f(z^i)^y \cdot (1 - f(z^i))^{1-y}$$

Al aplicar logaritmo natural se consigue la probabilidad logarítmica de la regresión logística:

$$LL(\theta) = \sum_{i=1}^{k} y\ log\ f(z)^y + (1 - y)\log\,(1 - f(z))$$

Como ya se tiene la función de verosimilitud, se debe de encontrar ahora los parámetros que maximicen dicha función, generalmente se realiza aplicando logaritmo natural a toda la función para después derivarla e igualarla a cero.

Condiciones para la regresión logística

Para un tener un buen funcionamiento del modelo se necesita que se cumplan todas las condiciones de este, sin embargo, en la mayoría de los casos, cuando se utilizan datos reales rara la vez se cumplen todas, pero eso no significa que el modelo no sea útil, de la misma manera se espera que, mientras más condiciones cumpla el modelo realice mejores predicciones.

No colinealidad o multicolinealidad

> En los modelos de regresión logística múltiple, los predictores deben ser independientes, no debe de haber colinealidad entre ellos. La colinealidad ocurre cuando un predictor está linealmente relacionado con uno o varios de los otros predictores del modelo (Amat, 2020).

Independencia

"Los valores de cada observación son independientes de los otros" (Amat, 2020).

Valores atípicos

Este tipo de valores son conocidos como outliers y hacen referencia a valores alejados de la zona en donde se encuentra la mayor parte de las muestras, cabe recalcar que estos valores no siempre son perjudiciales para el modelo.

2.2.2 Modelo de redes neuronales

Este modelo es uno de los más interesantes por la diversidad que puede tener en su composición y por lo general suele ser de los que mayor efectividad tienen en la clasificación de datos y algunos autores han comentado que:

> Es un algoritmo de aprendizaje supervisado que aprende una función $f(\cdot): R^m \rightarrow R^o$ mediante el entrenamiento en un conjunto de datos, donde *m* es el número de dimensiones para la

entrada y *o* es el número de dimensiones para la salida. Dado un conjunto de características $X = x_1, x_2, \ldots, x_m$ y un objetivo *y*, puede aprender un aproximador de función no lineal para clasificación o regresión (Scikit Learn, 2020).

De primera instancia este modelo, se puede expresar como una sola neurona, la cual está integrada por:

Un conjunto de entradas: es el conjunto de entrenamiento, donde cada observación X_j multiplica a sus respectivos pesos W_j, con *j* = 1, 2, …,n dichos valores se suman y se añade un término adicional que se llama bis, el cual es una constante, cabe mencionar que el peso de cada variable es desconocido, por lo tanto se debe de estimar.

De lo anterior tenemos que las entradas se pueden ver como: $x_1 w_1 + x_2 w_2 + \cdots + x_n w_n + b$.

Una función de propagación o activación: es una función que transmite la información generada por la combinación lineal de las entradas y los pesos, se denomina como $f_i(x_1 w_1 + x_2 w_2 + \cdots + x_n w_n + b)$ la cual representa la salida de la neurona, esta función va aplicada a todas las neuronas posteriores a las de la entrada, cabe recalcar que no necesariamente debe de ser la misma función para todas las neuronas, por ejemplo en este caso como estamos utilizando redes neuronales para hacer una clasificación binaria, la última neurona debe de estar asociada con una función de activación que tenga codominio en A, donde $A = \{0,1\}$.

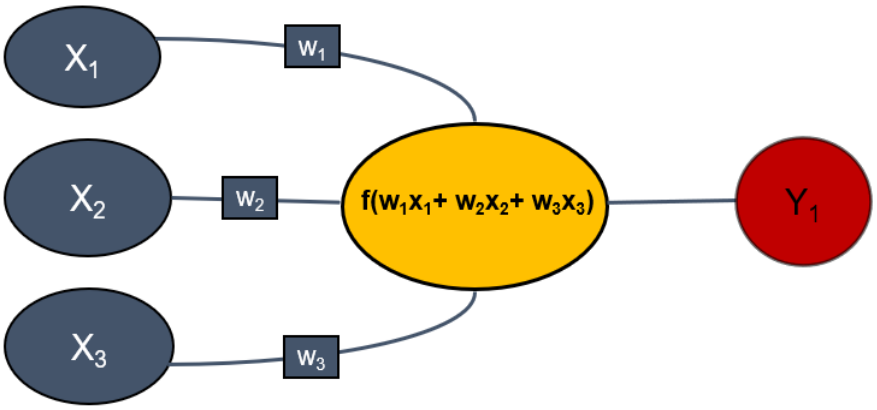


Figura 9. Esquema de redes neuronales.

De la Figura 9 se aprecia que el conjunto de entradas (óvalos azules) se conectan con sus respectivos pesos (cuadrados azules) y se transforman por medio de la función de activación (óvalo amarillo) de la cual se obtiene el resultado final (círculo rojo) por medio de otra función de activación. De lo anterior tenemos que el conjunto de entradas hace referencia al conjunto de datos de entrenamiento, el peso (W_i) debe de ser estimado, ya que se desconoce el valor real y las neuronas posteriores son predefinidas por medio de las funciones de activación que el analista haya asignado.

A continuación, se presentan las funciones de activación que generalmente se usan para las redes neuronales:

Función escalonada: Es una función definida a trozos, donde el analista asigna los valores de salida y los umbrales. A continuación, se presenta un ejemplo de la regla de correspondencia de esta función:

$$f(x) = \begin{cases} 2 & \text{si } x < X_0 \\ 4 & \text{si } x \geq X_0 \end{cases}$$

Función sigmoide: Esta función es la que se ocupa para realizar el modelo de regresión logit, para que todos los valores arrojados se encuentren dentro del intervalo [0,1], una característica de esta función es que los valores grandes se saturan cerca de 1 y los chicos cerca de 0, lo que ocasiona que la gráfica tenga una apariencia de una “S”. A continuación, se presenta la regla de correspondencia de esta función:

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

Función rectificada lineal: Esta función hace referencia al máximo valor entre el 0 y la suma ponderada, se utiliza principalmente cuando los valores negativos no tienen relevancia en el

modelo y por lo tanto, todos se acumulan en 0. A continuación, se presenta la regla de correspondencia de esta función:

$$f(x) = \begin{cases} 0 & si\ x<0 \\ x & si\ x\geq 0 \end{cases}$$

Función TANH: Esta función es muy parecida a la sigmoide, ya que toma prácticamente la misma forma de “S”, saturando los máximos valores en 1, con la diferencia de que los mínimos valores se acumulan en -1. A continuación, se presenta la regla de correspondencia de esta función:

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

El problema que se encuentra con solo separar los datos de esta manera es que la frontera no deja de ser una curva con respecto de la intersección de la figura geométrica con el plano. Para solucionar este problema se encadenan varias neuronas al mismo tiempo como se muestra a continuación:

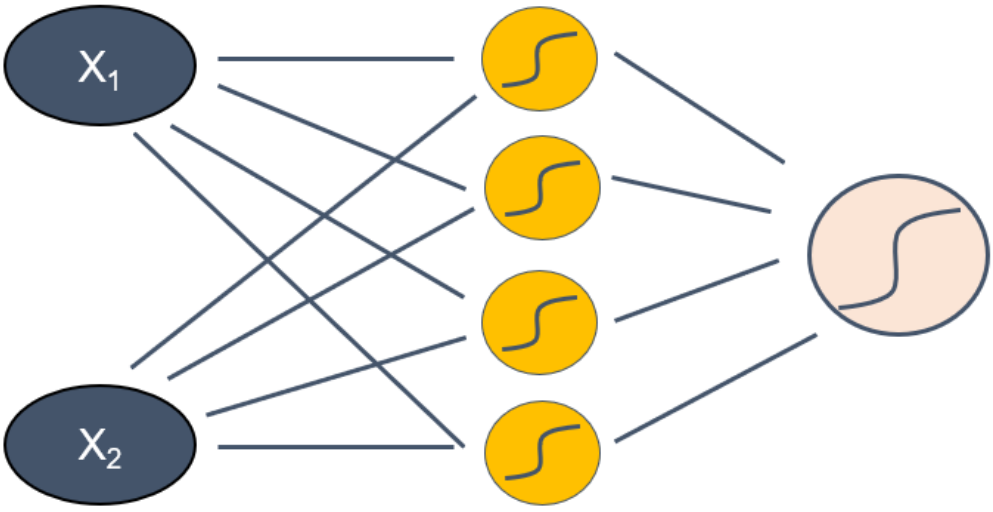


Figura 10. Esquema de redes neuronales

En la Figura 10 se añadieron más neuronas (círculos amarillos) para tener un procesamiento de los datos más complejo y que se muestre más flexible con respecto a ellos, cabe mencionar que no necesariamente implica que se obtengan mejores resultados ya que esto depende del tipo de datos que se estén evaluando.

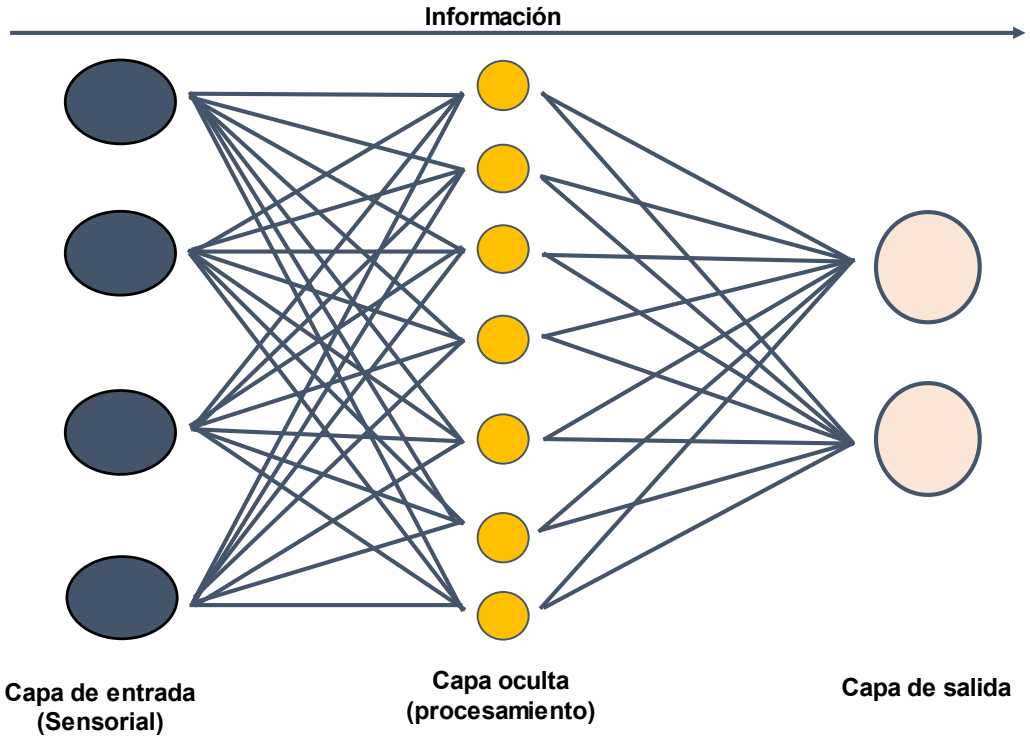


Figura 11. Esquema del modelo de redes neuronales

En la Figura 11 se presentó el esquema de una red neuronal, donde cada circulo de color hace referencia a una neurona y las cuales suelen agruparse en unidades estructurales (en la figura anterior hace referencia el color a cada unidad estructural) usualmente conocidas como capas y el conjunto de una o más capas es conocido como red neuronal. Argumentaron Larranaga, Inza, & Moujahid sobre las capas neuronales:

> Se distinguen tres tipos de capas: de entrada, de salida y ocultas. Una capa de entrada, también denominada sensorial, está compuesta por neuronas que reciben datos o señales procedentes del entorno. Una capa de salida se compone de neuronas que proporcionan la respuesta de la red neuronal. Una capa oculta no tiene una conexión directa con el entorno… este tipo de capa oculta proporciona grados de libertad a la red neuronal.

2.2.3 Árbol de decisión

Este modelo es conocido como árbol de decisión por la manera en que se desarrolla, parte de la condición inicial (raíz), de la cual surgen más condiciones (ramas). En la figura 12 se aprecia la estructura de este modelo, el nodo raíz se define como la condición 1 y es la condición más significativa, que es de donde parte el árbol, a cada nodo se le asigna una condición en donde, si se cumple se va hacia la rama en donde dice “si” (generalmente está a la izquierda del nodo), de lo contrario se va hacia donde dice “no”.

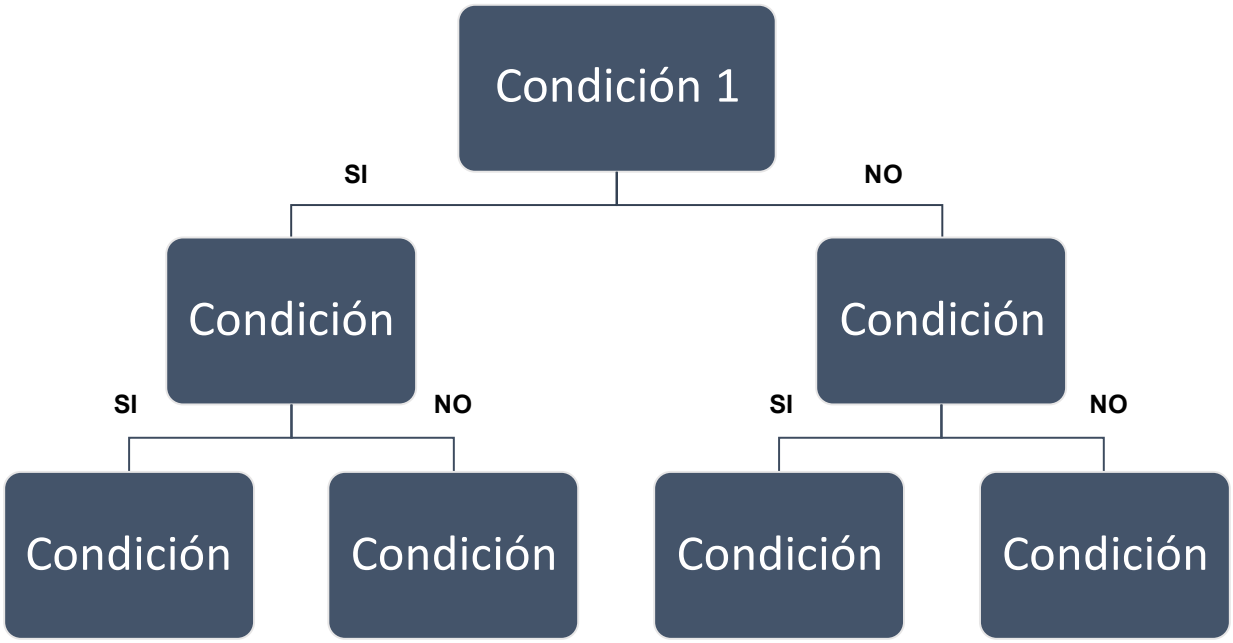


Figura 12. Esquema del Árbol de decisión

Scikit-Learn (2020) comentó la forma en que trabaja este modelo: “Un árbol puede verse como una aproximación constante por partes”, lo comentado anteriormente es porque se llega al resultado final por medio de condiciones, las cuales se complementan.

De la Figura 12 se aprecia que en la parte inferior es en donde mayor número de condiciones existen, pero en este caso como se utilizará como un clasificador binario, sin importar el número de condiciones solo puede arrojar dos resultados posibles.

Si tenemos que $X_1, X_2, \dots, X_p$ son las variables independientes y $R_1, R_2, \dots, R_j$ las regiones en donde se partirá el plano, entonces según James et al. (2013), el proceso para armar el árbol de decisión es:

1) Dividimos el espacio del predictor, es decir, el conjunto de valores posibles para $X_1, X_2, \dots, X_p$ en *J* regiones distintas y no superpuestas $R_1, R_2, \dots, R_j$.
2) Para cada observación que cae en la región R_j, hacemos la misma predicción, que es simplemente la media de los valores de respuesta para las observaciones de entrenamiento en R_j.

De primera instancia se necesita segmentar al conjunto de datos en regiones y realizar el promedio de todos los puntos que se encuentren en cada región y de esta manera estipular la condición de esa región, por ejemplo, si el promedio de una región es 10, tendríamos que si $x > 10$ entonces se le asigna un nodo, en el caso contrario se le asigna otro nodo.

De lo anterior, para segmentar al conjunto de datos en *J* regiones, se puede hacer de forma arbitraria, pero para tener una mejor clasificación suelen utilizarse dos técnicas.

La primera es el índice de Gini;

$$G = \sum_{k=1}^{K} \hat{P}_{mk}(1 - \hat{P}_{mk}) =$$

Donde $\hat{P}_{mk}$ es la proporción de las observaciones de entrenamiento en la región *m* que son de la k-ésima clase. James et al. (2013) argumentó sobre esta métrica aplicada a los árboles de decisión;

> No es difícil ver que el índice de Gini toma un valor pequeño si todos los $\hat{P}_{mk}$ están cerca de cero o uno. Por esta razón, el índice de Gini se conoce como una medida de la pureza del nodo: un valor pequeño indica que un nodo contiene predominantemente observaciones de una sola clase.

La segunda técnica es *cross-entropy;*

$$D = -\sum_{k=1}^{K} \hat{P}_{mk} Log \hat{P}_{mk}$$

Donde se espera que los valores de estas métricas sean lo más homogéneo posible, es decir que se encuentren cerca del cero, al igual que el índice de Gini.

2.2.4 Análisis discriminante lineal

El análisis discriminante lineal es uno de los principales clasificadores que existen por su gran eficacia y sencillez al momento de elaborarlo, este modelo es mejor conocido por LDA, que significa Lineal Discriminant Analysis y algunos autores han comentado que:

> El Análisis Discriminante (Discriminant Analysis, DA) es una de las técnicas de análisis multivariante más conocidas cuyo objetivo es encontrar la combinación lineal (o cuadrática) de las variables independientes (también llamadas variables de clasificación en las que suponemos que se diferencian los grupos) que mejor permitan diferenciar (discriminar) a los grupos (clases), por medio de una función discriminante, la cual podrá ser utilizada para clasificar nuevos casos (Henríquez, 2014).

El método de análisis discriminante se divide en dos, cuando este es lineal (LDA) y cuando es cuadrático (QDA), pero en este trabajo solo se tomará en cuenta el lineal. La variable dependiente es categórica, haciendo referencia a los grupos predefinidos, los cuales son las categorías en las que se quiere segmentar las observaciones y por otro lado tenemos que las variables independientes que son los atributos de cada observación.

Para poder presentar el modelo se definen dos variables; sea k el número de categorías en las que se quiere segmentar al conjunto de datos y π_k, la probabilidad de que una observación dada esté asociada con la k- ésima categoría de la variable dependiente *Y*.

Entonces definamos que: $f_k(x) = P(X = x|Y = k)$ como la función de probabilidad de densidad para *X*, dado que proviene de la categoría *k*.

Recordamos que el teorema de bayes nos dice que:

$$P(Y = k|X) = \frac{P(X|Y = k) \cdot P(Y = k)}{P(X)} = \frac{P(X|Y = k) \cdot P(Y = k)}{\sum_l P(X|Y = l) \cdot P(Y = l)}$$

Entonces según James et al. (2013), por el teorema anterior se tiene que:

$$P(Y = k|X = x) = \frac{\pi_k \cdot f_k(x)}{\sum_{l=1}^{K} \pi_l \cdot f_l(x)}$$

Donde:

- $\hat{\pi}_k = {}^{N_k}/_{N}$

$donde\ N_k\ es\ el\ número\ de\ observaciones\ de\ la\ clase\ k$ y $N\ el\ número\ total\ de\ observaciones.$

Si suponemos que las variables predictoras son independientes e idénticamente distribuidas con $E[x_i] = \mu$ y $Var[x_i] = \sigma^2$, entonces por medio del teorema del límite central podemos comentar que $f_k(x)$ se aproxima bien a una distribución normal.

$$f_k(x) \approx \frac{1}{\sqrt{2\pi\sigma_k}} exp\left(-\frac{1}{2\sigma_k^2}(1 - \mu_k)^2\right)$$

Por lo tanto, la probabilidad de que la variable dependiente sea *k*, dado *X* es:

$$P(Y = k|X) = \frac{\pi_k \ \frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{1}{2\sigma^2}(1 - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{1}{2\sigma^2}(1 - \mu_k)^2\right)}$$

Condiciones para LDA

Para que este modelo pueda tener una mayor efectividad al realizar las clasificaciones las variables deben de cumplir ciertas características:

- Entradas numéricas, es decir que todos los valores de la base de datos sean numéricos.
- Todas las variables deben de tener distribución normal.
- Matriz de covarianza común con las variables independientes.
- Variables no correlacionadas.

De la misma manera, siempre se debe de tratar que los datos cumplan con todas estas características para que el modelo funcione de la mejor manera, pero incluso si no se cumplen todas suele dar un buen rendimiento.

Ahora que se toma en cuenta el cálculo de este modelo con el clasificador de regresión logit se tiene que son muy parecidos, ya que ambos son clasificadores lineales, pero tienen una gran diferencia, las cuales son sus condiciones, de las cuales argumentó James et al. (2013):

> LDA asume que las observaciones se extraen de una distribución gaussiana con una matriz de covarianza común en cada clase, por lo que puede proporcionar algunas mejoras sobre la regresión logística cuando esta suposición es aproximadamente válida. Por el contrario, la regresión logística puede superar a LDA si no se cumplen estos supuestos gaussianos.

En la cita anterior, la distribución gaussiana hace referencia a la distribución normal.

2.2.5 Máquina de vectores de soporte

El modelo de Máquina de vectores de soporte, conocido como Support Vector Machines (SVM), es un conjunto de algoritmos de aprendizaje supervisado el cual se utiliza para realizar regresiones o clasificar observaciones.

> SVM es un clasificador lineal en el que busca minimizar el error en relación con el conjunto de muestras de entrenamiento (riesgo empírico) y el error en relación al conjunto de muestras de prueba (riesgo en generalización) (Andreola & Haertel, 2010).

Para entender este modelo, se presenta a continuación la definición de hiperplano y se introducen algunos conceptos sobre la separación de hiperplanos.

Según Andreola & Haertel (2010) el hiperplano es:

> En un espacio p-dimensional, un hiperplano es un subespacio afín plano de dimensión p - 1.1. Por ejemplo, en dos dimensiones, un hiperplano es un subespacio plano unidimensional, en otras palabras, una línea. En tres dimensiones, un hiperplano es un subespacio bidimensional

plano, es decir, un plano. En p> 3 dimensiones, puede ser difícil visualizar un hiperplano, pero la noción de un subespacio plano (p - 1) dimensional todavía se aplica.

Según James (2013), la definición de hiperplano en dos dimensiones se presenta de la siguiente forma:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

Tenemos que los parámetros de la ecuación son β_0, β_1 y β_2 para cualquier $X = (X_1, X_2)^t$ donde se cumpla la condición anterior entonces tenemos que dichos valores serian un punto en el hiperplano.

Se tiene que dado el número de las variables que son X_1 y X_2 el hiperplano anterior se refleja que es un subespacio unidimensional plano, o sea una línea. Notar que de la ecuación anterior se puede extender al modelo p-dimensional, como se presenta a continuación:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$$

Donde X nuevamente cumple la propiedad antes mencionada $X=(X_1,\ldots, X_p)^T$ .Los puntos donde X no cumple la propiedad anterior es donde son diferente de cero, por lo tanto los valores son positivos o negativos.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p > 0$$

En la ecuación anterior se presenta la ecuación tal que los valores de X son mayores que cero, por otro lado, se presenta a continuación la expresión cuando son menores a cero.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p < 0$$

Ahora que ya comprendemos el concepto de hiperplano, se puede dar una definición más exacta sobre la máquina de vectores de soporte y Henriquez (2014) lo argumentó de la siguiente manera;

> Las Máquinas de Vectores de Soporte (Support Vector Machine, SVM) se emplean tanto en aplicaciones de clasificación como en regresión y se basa en la determinación del hiperplano que da lugar a la máxima distancia de separación entre los vectores transformados … esta distancia de separación se obtiene mediante la construcción de dos hiperplanos paralelos al

hiperplano de separación óptimo, localizados a ambos lados del mismo y que contengan al menos a uno de los vectores transformados, denominados Vectores de Soporte … se asume que cuanto mayor sea esta distancia, mejor es la capacidad de generalización del clasificador.

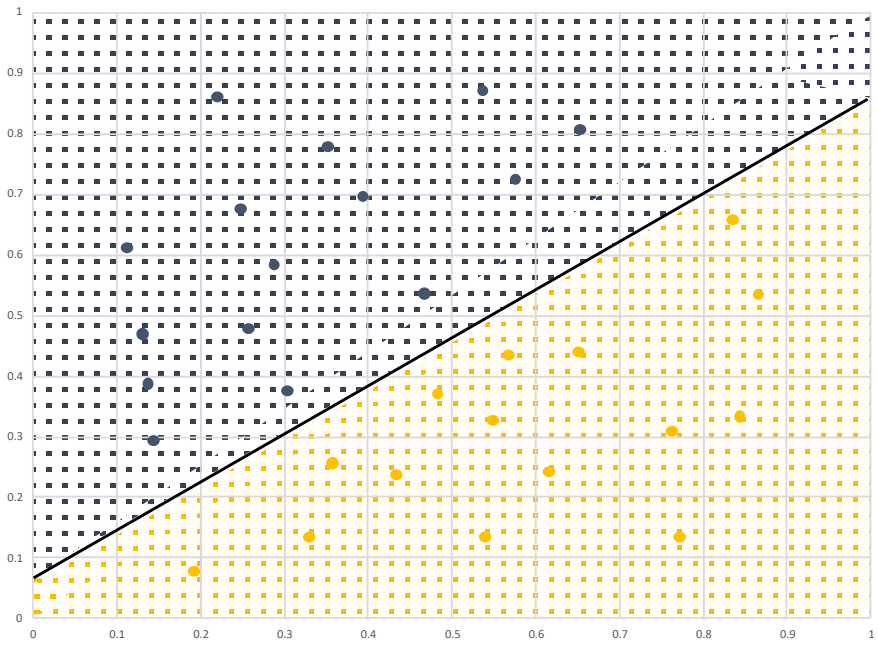


Figura 13. *Gráfico de SVM.*

En la Figura 13 se aprecia como un hiperplano separa los tipos de observaciones que existen, tomando en cuenta que los puntos que caigan en la zona azul, que se encuentra en la parte superior de la izquierda, se etiqueten como $y_i = 1$, por el otro lado, los puntos que caigan en la zona amarilla se etiquetan con $y_i = -1$. Entonces tenemos que:

$$\beta_0 + \beta_1 Xi_1 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} > 0 \text{ si } y_i = 1$$

$$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} < 0 \; si \; y_i = -1$$

Si existe la separación del hiperplano, entonces podemos construir un clasificador muy intuitivo, donde si tenemos que, por medio de las variables explicativas, la ecuación arroja un resultado menor a cero, se le asignaría el -1 a y, de lo contrario sería el 1.

Antes de obtener los coeficientes de este modelo, se presentan a continuación las variables que nos ayudarán para la obtención de estos.

Vectores de soporte. Hace referencia a las observaciones que están más cercanas al hiperplano que segmenta los tipos de observaciones.

Sea $M = \frac{1}{\|\beta\|}$, donde β es la separación que existe entre el hiperplano y los vectores de soporte.

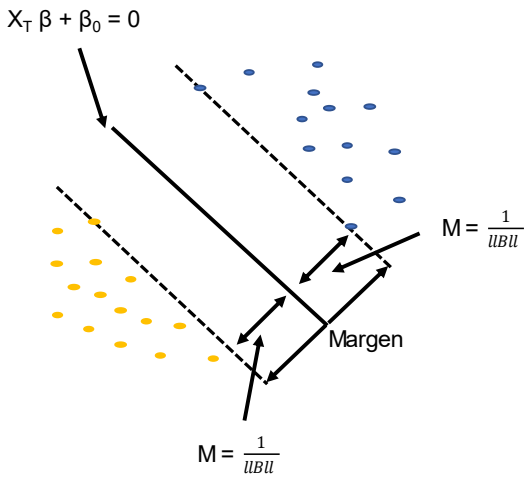


Figura 14. Representación de la variable M.

Para conseguir los coeficientes $\beta_0, \beta_1, \dots, \beta_p$ utilizaremos el conjunto de entrenamiento de *n* observaciones $x_1, \dots, x_n \in R^p$, con sus respectivas categorías $y_1, \dots, y_n \in \{-1,1\}$. El objetivo principal es conseguir el máximo margen de separación del hiperplano con respecto de los vectores de soporte, por lo tanto, tenemos que:

$$\underset{\beta_0, \beta_1, \dots, \beta_p}{\text{Maximizar M}} \quad \text{Sujeto a} \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M \quad \forall\, i = 1, \dots, n.$$

Por medio de la última restricción podemos asegurarnos que cada observación estará en el lado correcto del hiperplano, siempre que *M* sea positivo. Tomando en cuenta ambas restricciones es fácil ver que la distancia perpendicular que existe entre el hiperplano y la i-esima observación está dada por:

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$$

Ahora que sabemos a lo que hace referencia la función anterior, tenemos como objetivo encontrar la máxima distancia posible por medio de la variable *M*.

Las principales acciones que realiza SVM, según Henriquez (2014) son:

Transformación de los datos o vectores de características de entrada a un espacio de mayor dimensión a través de una función núcleo o kernel *K*.

Cálculo del hiperplano ´optimo que maximiza la distancia entre las clases consideradas. Si los datos son linealmente separables, el hiperplano obtenido maximiza el margen de separación, a la vez que minimiza la función de penalización que considera las clasificaciones incorrectas.

Tomando en cuenta la complejidad que existe en separar las observaciones, se puede encontrar que no sean linealmente separables o que existe un nivel de ruido. Dado lo anterior se pueden tomar diversos tipos de SVM para solucionar los problemas: (1) SVM lineal con margen máximo, (2) SVM con margen blando o (3) SVM para la clasificación no lineal.

1.- SVM lineal con margen máximo, se ocupa cuando existe la manera de separar linealmente lo datos.

2.- SVM con margen blando, se emplea cuando no existe un hiperplano optimo que pueda separar las observaciones de entrenamiento, para obtener el mejor hiperplano se separación se permite que algunas observaciones estén en el lado incorrecto del margen.

3.- SVM para la clasificación no lineal, es una variación de la SVM en donde el hiperplano que divide a las muestras de entrenamiento es una curva en lugar de ser una recta, por lo tanto, es

más flexible y ocasiona que este SVM pueda tener un hiperplano que pueda separar a la mayoría de los datos.

2.3 Modelos de aprendizaje no supervisado

2.3.1 Modelo de k- medias

Este modelo es completamente diferente a los ya demostrados anteriormente, ya que es de aprendizaje no supervisado y se piensa que:

> El algoritmo de K-Medias agrupa los datos tratando de separar muestras en n grupos de igual varianza, minimizando un criterio conocido como inercia o suma de cuadrados dentro del grupo. Este algoritmo requiere que se especifique el número de clústeres. Se adapta bien a una gran cantidad de muestras y se ha utilizado en una amplia gama de áreas de aplicación en muchos campos diferentes (scikit-learn, 2020).

En este modelo primero se especifica el número de categorías (*K*) en las que se desea dividir al conjunto de datos, para que el modelo asigne cada observación exactamente a uno de los k grupos. Sea n el número total de observaciones a clasificar y $C_1, \dots, C_k$ los conjuntos que contienen los índices de las observaciones en cada grupo.

Según James et al. (2013), los conjuntos $C_1, \dots, C_k$ deben de satisfacer dos propiedades:

1.- $C_1 \cup C_2 \cup \dots \cup C_k = \{1, \dots, n\}$.

2.- $C_p \cap C_s = \emptyset$ para todo p diferente de s.

De la primera propiedad se entiende que la unión de todos los subgrupos debe de ser igual al conjunto de datos inicial, con las n observaciones antes mencionada. La segunda propiedad nos dice que la intersección entre dos grupos es igual a vacío, por lo tanto, ninguna observación estará en más de una categoría.

Como se mencionó anteriormente, la clave del modelo es que, dentro de cada agrupamiento, la variación entre las observaciones sea la mínima posible. Argumentó James et al. (2013), sobre el modelo de k- medias;

> La variación dentro del cluster para el conglomerado C_k es una medida $W(C_k)$ de la cantidad en la que las observaciones dentro de un conglomerado difieren entre sí. Por eso queremos resolver el problema.

$$\underset{C_1,\dots,C_k}{\text{Minimizar}} \left\{\sum_{k=1}^{K} W(C_k)\right\}$$

En resumen, tenemos que, $W(C_k)$ es la variación que existe en cada cluster y dado que se suman los *k* cluster, tendríamos la suma de la variación de todos los cluster y como se minimiza toda esta función, se entiende con esta fórmula que se busca minimizar la variación en todos los cluster, así como lo mencionó el autor en el párrafo anterior.

Para poder resolver este problema, se necesita encontrar la forma de calcular la variación en cada cluster, existe más de una manera para definir este concepto, pero generalmente se realiza por medio de la distancia euclidiana, que se define como;

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,\acute{\imath}\in C_k} \sum_{j=1}^{P} \left(x_{ij} - x_{\acute{\imath}j}\right)^2$$

Donde C_k denota el número de observaciones en cada k cluster. En resumen, tenemos que para realizar este método se tiene la siguiente formula y generalmente se resuelve por medio de iteraciones:

$$\underset{C_1,\dots,C_k}{\text{Minimizar}} \left\{\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,\acute{\imath}\in C_k} \sum_{j=1}^{P} \left(x_{ij} - x_{\acute{\imath}j}\right)^2\right\}$$

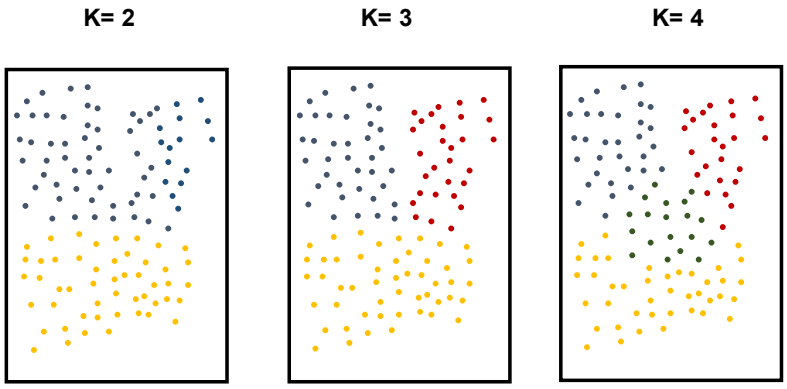


Figura 15. Ejemplo ilustrativo del modelo K-Medias.

En la Figura 15 se mostró un conjunto de datos simulados con 150 observaciones en un espacio bidimensional. Arriba de cada panel está la variable “k”, que hace referencia al número de clusters en los que se clasificó al conjunto de datos, tomando en cuenta que a cada observación se le asignó un color ya sea, amarillo, azul, verde o rosa, dependiendo al cluster al que pertenece.

2.4 Métricas para modelos

Ahora que ya se tiene el conocimiento de la elaboración de un clasificador se debe de encontrar la forma de compararlos entre sí para poder elegir el óptimo por medio de pruebas estadísticas. Existe una gran variedad de pruebas para analizar el poder predictivo que tiene cada clasificador pero en este trabajo sólo se hizo énfasis a los métodos más usados para los clasificadores de créditos.

2.4.1 Curva de característica operativa del receptor (ROC)

La curva de característica operativa del receptor, generalmente conocida como la curva ROC es, una de las pruebas más importantes para verificar el nivel de predicción de los modelos de clasificación y Valle argumentó sobre ello:

> La curva ROC es una herramienta estadística utilizada en el análisis de la clasificar la capacidad discriminante de una prueba diagnóstica dicotómica. Es decir, una prueba, basada en una variable de decisión, cuyo objetivo es clasificar a los individuos de una población en dos grupos: uno que presente un evento de interés y otro que no.

En resumen, tenemos que la curva ROC, nos dice que tanto un clasificador puede distinguir entre dos tipos de observaciones, teniendo en cuenta que, mientras mayor precisión tenga el modelo, más útil será, ya que este las podrá distinguir de mejor manera.

A continuación, se presenta la matriz de confusión en la Figura 17 para poder explicar algunas variables:

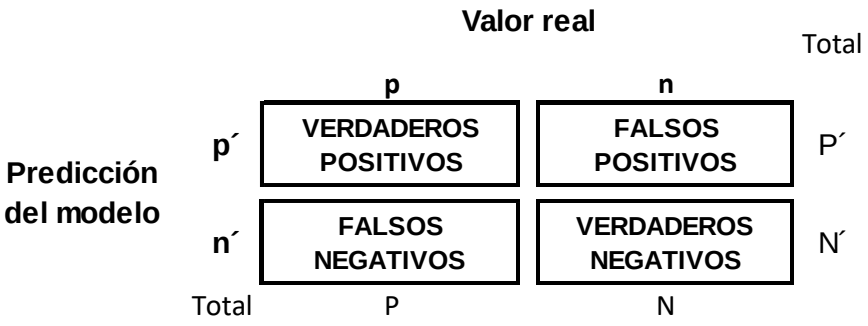


Figura 16. Matriz de confusión.

Sea un ejemplo donde el modelo arroje el número 1 para cuando el individuo será moroso y 0 cuando este no lo sea.

Verdaderos positivos (**VP**): Hace referencia al evento donde el individuo es moroso y se clasificó como moroso.

Verdaderos negativos (**VN**): Este individuo no es moroso y se clasificó como no moroso.

Falsos positivos (**FP**): El individuo no es moroso y se clasificó como moroso (Error tipo 1)

Falsos negativos (**FN**): El individuo es moroso y se clasifico como no moroso (Error tipo 2)

Cabe recalcar que, para obtener las siguientes métricas se utilizó el axioma de Kolmogorov de "La probabilidad de un suceso es el número de casos favorables entre el número de casos posibles".

Sensibilidad o razón de verdaderos positivos (VPR)

Hace referencia a que un individuo que, si va a pagar el crédito, el modelo lo clasifique como bueno.

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

Ratio o razón de falsos positivos (FPR)

Es la razón de falsas alarmas, el número de no morosos falsos con respecto de los no morosos reportados.

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

Exactitud (ACC)

Como el nombre lo dice, hace referencia al número de elecciones correctas con respecto del total.

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

Especificidad (SPC) o razón de verdaderos negativos

Hace referencia al número de no morosos que fueron elegidos correctamente con respecto de todos los no morosos reportados, cabe recalcar que esta métrica también se puede calcular como el complemento de FPR.

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

Valor predictivo positivo (PPV)

Calcula el número de individuos morosos y se clasificaron como tal, con respecto de todos los que se clasificaron como morosos.

$$PPV = \frac{VP}{VP + FP}$$

Valor predictivo negativo (NPV)

Hace referencia a los que se clasificaron correctamente como no morosos con respecto de todos los que se clasificaron como no morosos.

$$NPV = \frac{VN}{VN + FN}$$

Ratio o razón de falsos descubrimientos (FDR)

Hace referencia a los individuos con los que se cometió el error tipo 1 con respecto de los que se clasificaron como morosos.

$$FDR = \frac{FP}{FP + VP}$$

Cabe recalcar que todas estas métricas son muy importantes para los modelos, dado que, no solo podemos saber qué nivel de predicción tuvo, si no, también sirve para analizar si el clasificador tuvo más errores tipo 1 o errores tipo 2, un ejemplo de esto es cuando la institución financiera está en una etapa conservadora, en este caso se preferiría un modelo que cometa más errores tipo 1 que 2 y así tener muchos más clientes no morosos que morosos.

La prueba ROC tiene una amplia gama de estadísticos para evaluar a los clasificadores, sin embargo, en este trabajo utilizaremos solo uno, el cual es el más importante, el área bajo la curva ROC, llamada comúnmente AUC y Valle argumentó sobre ello:

> Una primera forma de calcularla sería representando la curva ROC, o una estimación de ella, y obtener el porcentaje de área del cuadrado [0, 1] × [0, 1] que encierra bajo ella. Veremos a continuación que no siempre será necesario tener la curva para obtener esta área, sino, que a partir de los datos podremos estimarla directamente.

Existe más de una manera para realizar el cálculo de la curva ROC, pero la más conocida es por el cálculo de la regla trapezoidal el cual es la suma de áreas de trapecios puesto que tendremos una curva en forma de escalera, su fórmula es:

$$AUC = \sum_{t=1}^{T} \frac{1}{2}(FFP_t - FFP_{t-1}) \cdot (FVP_t + FVP_{t-1})$$

Donde FFP_t y FVP_t son las fracciones de falsos positivos y fracciones de verdaderos positivos respectivamente calculadas para cada t = 1, ..., T puntos de corte.

En la Figura 18**,** se aprecia el grafico de la curva ROC con tres tipos de niveles de predicción, cuando tenemos que AUC=1, existe un nivel de discriminación excelente mientras que, por el otro lado, cuando AUC=0.5, el modelo no tiene un nivel de discriminación significante.

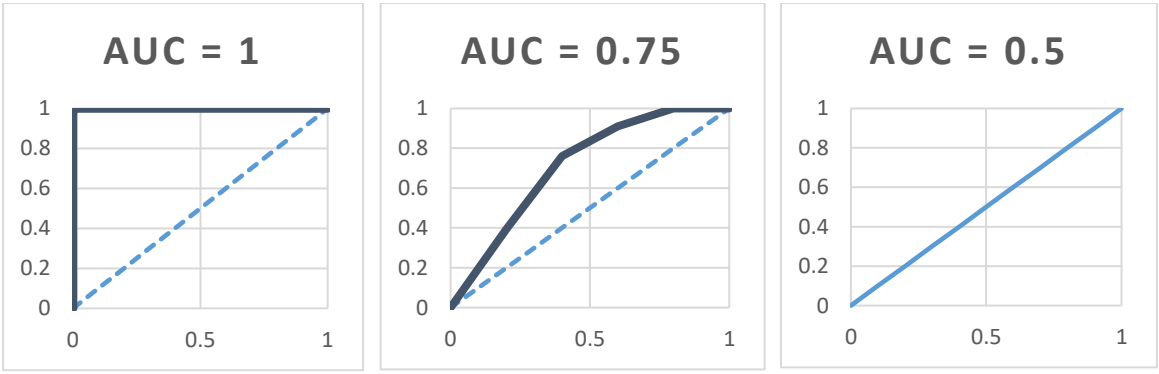


Figura 17. Gráfico de la curva ROC.

Tenemos que la excelencia en la curva es cuando existe AUC =1 y el peor de los casos con AUC= 0.5, pero ¿cómo podemos saber cuándo el modelo es relativamente bueno y cuando malo? A continuación, en la Tabla 1, se presenta la tabla con los diferentes rangos sobre los niveles de predicción.

Baja exactitud	[0.5,0-7)
útiles para algunos propósitos	[0.7,0.9)
Exactitud alta	[0.9,1]

Tabla 1 *Niveles de exactitud para la métrica AUC-ROC.*

2.4.2 Índice de Gini

Una de las ventajas que tiene esta prueba es su elaboración ya que es rápida y de fácil comprensión, aparte de ser muy relevante y Lizárraga comentó sobre ello:

> El indicador más utilizado y aceptado para cuantificar los niveles de desigualdad de ingresos es el coeficiente de concentración de Gini, que toma valores entre 0, cuando existe completa igualdad en la distribución del ingreso, y 1, en caso de completa desigualdad.

La curva de Lorenz hace referencia a la proporción perfecta de riqueza que existe por persona, por ejemplo, el 30% de la población se le asigna el 30% de riqueza y mientras más alejada esté una economía de esta curva se dice que existe una mayor desigualdad de ingresos.
Esta prueba nosotros la usaremos, pero en lugar de tener el nivel de riqueza y la curva de Lorentz, utilizaremos la distribución de buenos y malos tomando en cuenta que el valor mínimo aceptable suele ser .30 y se tiene que, mientras mayor sea el numero arrojado, existirá una mayor discriminación entre las distribuciones, por lo tanto, se espera que un buen clasificador se encuentre lo más alejado del cero.

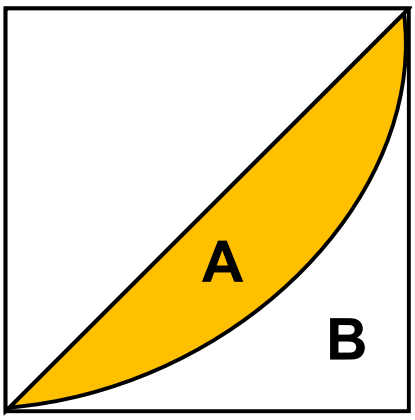


Figura 18. Representación gráfica del cálculo del índice de Gini.

Medina (2001) comentó sobre esta métrica que: “existen diversas formas de derivar la expresión algebraica que se usa para su cálculo y también es posible deducirlo desarrollando un procedimiento geométrico a partir de la curva de Lorenz”. Es decir, se quiere calcular la proporción del área A con respecto del área A y B, por lo tanto, la formula seria $a/(a+b)$, lo cual se ve representado en la Figura 19.

Según Lizárraga “el cálculo del índice de Gini se lleva a cabo de diversas formas, aunque la más extendida es la fórmula de Brown”

$$CG = \left| 1 - \sum_{k=1}^{n-1} (X_{k+1} - X_k)(Y_{k+1} + Y_k) \right|$$

Donde

CG: Coeficiente de Gini

X_k: Proporción acumulada de la variable de clientes buenos del percentil k

Y_k: Proporción acumulada de la variable de clientes malos del percentil k

2.4.3 *Information value*

La métrica Information Value es una de las pruebas más relevantes cuando se trata de problemas binarios y en específico cuando se trata de clasificar a los tipos de clientes entre buenos y malos, los cuales hacen referencia a los que sí pagan y a los morosos respectivamente.

Esta métrica va de la mano de la variable llamada “el peso de la evidencia” del inglés Weight Of Evidence (WOE) y argumentó Radecic (2019) que: “El peso de la evidencia indica el poder predictivo de una variable independiente en relación con la variable dependiente”. El WOE es el logaritmo de la proporción de la distribución de los usuarios buenos con respecto de los malos:

$$WOE_i = ln\left[\frac{\%\ clientes\ buenos_i}{\%\ clientes\ malos_i}\right] \text{ para i =1, … ,n}$$

Donde

$$\%\ clientes\ buenos_i = \frac{\#\ clientes\ buenos_i}{Total\ clientes\ buenos}$$

$$\%\ clientes\ malos_i = \frac{\#\ clientes\ malos_i}{Total\ clientes\ malos}$$

Esta prueba mide el área entre la distribución de los clientes buenos y la distribución de los clientes morosos y se calcula de la siguiente forma:

$$IV = \sum(\%\ buenos_i - \%\ malos_i) \cdot WOE_i$$

Donde

IV: Information Value

$$\%\ buenos_i = \frac{\#\ clientes\ buenos\ de\ la\ seccion\ i}{Total\ de\ clientes\ buenos}$$

$$\%\ malos_i = \frac{\#\ clientes\ malos\ de\ la\ seccion\ i}{Total\ de\ clientes\ malos}$$

Entre menor sea IV existe una menor relación de la variable independiente con respecto de la dependiente, a continuación, en la Tabla 2 se presenta según Tibco el rango de los valores y su correspondiente interpretación:

Information Value	Poder predictivo
<0.02	No es útil
0.02 – 0.1	Débil
0.1-0.3	Medio
0.3 – 0.5	Fuerte
>0.5	Poder predictivo sospechoso

Tabla 2. *Valores de Information Value*.

2.4.4 Prueba Kolmogorov -Smirnov

La prueba Kolmogorov -Smirnov (K-S) es una prueba no paramétrica que se utiliza para comparar dos distribuciones y analizar si existe suficiente diferencia entre ellas o no. En este trabajo se planteará la hipótesis de no diferencia significativa en las distribuciones de buenos y malos clientes contra diferencia significativa de las mismas. En teoría tenemos que mide la máxima separación vertical entre la distribución acumulada de buenos y malos:

$$D_n = máx|F_e - F_o|$$

Donde

D_n: Estadístico de la prueba K-S

F_e: Distribución acumulada de buenos

F_o: Distribución acumulada de malos

Recuperando lo anterior podemos decir que a cada clasificador se le tiene que realizar las pruebas pertinentes con el objetivo de seleccionar el modelo que mejor se ajuste a los datos evaluados y con base a esto conocer el nivel de predicción que se tenga del mismo.

Capítulo 3.- Construcción del modelo de score

En este capítulo se muestran las descripciones de las variables del conjunto de datos utilizados para la elaboración de los clasificadores con su respectivo análisis a partir de gráficas y tablas, tomando en cuenta que se crearon tres variables a partir del conjunto de datos ya proporcionado, esto con la finalidad de tener un mejor rendimiento en los clasificadores.

Por último, se presenta la construcción de cada clasificador y los resultados obtenidos a partir de las técnicas de validación, de la misma manera se elaboró un análisis para poder concluir cual fue el mejor modelo elaborado para el conjunto de datos dado.

3.1 Descripción de la base de datos

La base de datos a evaluar contiene características de los usuarios que tuvieron tarjeta de crédito de un banco en Taiwán, en el periodo de abril de 2005 a septiembre de 2005. El conjunto de datos fue presentado por I-Cheng Yeh y publicado por el centro de machine learning y sistemas de inteligencia de la universidad de California, Irvine.

La base de datos está compuesta por 30,000 observaciones con 23 variables independientes y 1 dependiente, a continuación, en la Tabla 3 se presenta la descripción de cada una de las variables tomando en cuenta que los valores monetarios se encuentran en dólares:

Nombre	Descripción
Límite de crédito	Incluye crédito individual y familiar (suplementario).
Sexo	1= masculino, 2 = femenino.
Educación	1= escuela de posgrado, 2= universidad, 3= escuela secundaria, 4 = otros.

Estado civil	1 = casado, 2 = soltero, 3 = otros.
Edad	Edad en años.
Pago 1	Estado de pago en septiembre de 2005, -1 = pago debidamente hecho, 1 = retraso en el pago por un mes, 2= retraso en el pago por dos meses, …, 8 = retraso en el pago por ocho meses, 9 = retraso en el pago por nueve meses o más.
Pago 2	Estado de pago en agosto de 2005 (misma escala a la de Pago 1).
Pago 3	Estado de pago en julio de 2005 (misma escala a la de Pago 1).
Pago 4	Estado de pago en junio de 2005 (misma escala a la de Pago 1).
Pago 5	Estado de pago en mayo de 2005 (misma escala a la de Pago 1).
Pago 6	Estado de pago en abril de 2005 (misma escala a la de Pago 1).
Monto del estado de cuenta 1	En septiembre del 2005.
Monto del estado de cuenta 2	En agosto del 2005.
Monto del estado de cuenta 3	En julio del 2005.
Monto del estado de cuenta 4	En junio del 2005.
Monto del estado de cuenta 5	En mayo del 2005.
Monto del estado de cuenta 6	En abril del 2005.
Monto del pago anterior 1	En septiembre del 2005.
Monto del pago anterior 2	En agosto del 2005.

Monto del pago anterior 3	En julio del 2005.
Monto del pago anterior 4	En junio del 2005.
Monto del pago anterior 5	En mayo del 2005.
Monto del pago anterior 6	En abril del 2005.
Impago en el siguiente mes	1 = sí, 0 = no.

Tabla 3. Nombre de las variables con su descripción.

Al observar la base de datos y sus características, se encontraron irregularidades y algunos valores en la base de datos que no son contemplados en su descripción, por lo tanto, se optó por realizar algunas modificaciones, las cuales son las siguientes:

- La variable **Educación** contiene valores del 0 al 6, pero la descripción solo contempla del 1 al 4, por lo tanto, no se tiene descripción sobre los valores 0, 5 y 6, entonces se agruparon junto con el valor 4, el cual hace referencia a “otros”.
- La variable **Estado Civil** contiene valores del 0 al 3, dado que el 0 en la descripción no se contempla este valor se tomó como si fuera el 3, el cual hace referencia a “otros”.
- Las variables **Pago i** con i = 1, 2, 3, 4, 5 y 6 contienen valores del -2 al 8, por lo tanto, se optó por agrupar el -2,-1 y 0 haciendo referencia a que el cliente hizo el pago debidamente hecho.
- Las variables **Monto del estado de cuenta i** con **i** = 1, 2, 3, 4, 5 y 6 contemplan valores mayores al límite de crédito otorgado, por lo tanto, dichos valores se acotaron a la variable **Límite de crédito**, de la misma manera existen valores negativos, los cuales se reemplazaron por ceros, haciendo referencia a que el usuario no le debe al banco.
- Las variables **Monto del pago anterior i** con **i** = 1, 2, 3, 4, 5 y 6 contiene valores mayores a la cantidad que se debe, por lo tanto, dichos valores se acotan a su deuda.

3.2 Análisis de las variables

En esta sección se hizo un análisis crítico de los datos a evaluar, obteniendo información de las variables que se utilizaron para realizar el clasificador. Este análisis se basa principalmente de estadística descriptiva y gráficos, los cuales ayudaron para encontrar su variabilidad, media y observar la forma en que se distribuyen.

Para poder presentar las gráficas de frecuencia de las variables continuas se tuvieron que transformar en discretas, tomando en cuenta que se optó únicamente por 11 intervalos y para obtener dichos intervalos se utilizó la siguiente formula $\frac{Rango}{intervalos}$.

Para la variable **Límite de crédito** en la Tabla 4 se encontró que la diferencia que existe entre el valor máximo de la variable y el mínimo es de 990,000, lo cual hace referencia a una varianza entre los valores de la variable relativamente grande y se puede ver reflejado en su desviación estándar. Al observar que la media está más cerca del mínimo que del máximo, nos podemos dar cuenta que la gráfica de los valores debe de estar inclinada hacia los valores chicos, es decir, el banco tiende a dar más líneas de crédito cercanas al mínimo que al máximo.

Mínimo	10,000.00	**Media**	167,484.32
Máximo	1,000,000.00	**Desviación estándar**	129,747.66
Rango	990,000.00	**Numero de Observaciones**	30,000.00

Tabla 4. *Valores de* ***Límite de crédito****.*

Para tener un mayor conocimiento de cómo se distribuye la variable **Límite de crédito***,* a continuación*,* en la Figura 20 se presenta su gráfica de frecuencia.

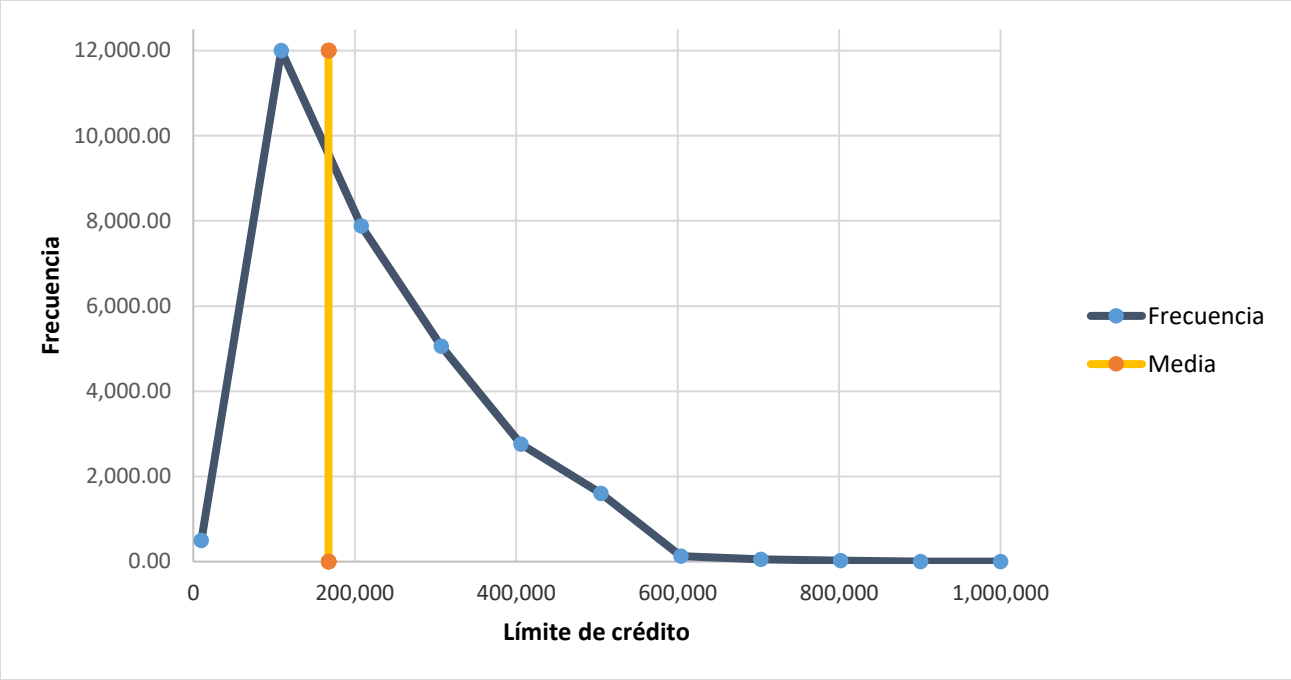


Figura 19. *Gráfica de **Límite de crédito**.*

En la Tabla 5 se observa que del 100% de las personas a las que se les aprobó la tarjeta de crédito, la mayoría son mujeres con un 60%.

Sexo	Frecuencia	Porcentaje
Hombres	11,888.	40%
Mujeres	18,112	60%
Total	30,000	100%

Tabla 5. *Valores de **Sexo**.*

De la Tabla 6, se analizó que a la mayoría de las personas que se les otorgó el crédito fueron las que tienen un posgrado o la universidad terminada, esto se debe a que regularmente cuando una persona tiene más estudios tiene un trabajo más estable y por lo mismo podrá pagar sus deudas.

Educación	Frecuencia	Porcentaje
Posgrado	10,585	35%
Universidad	14,030	47%
Secundaria	4,917	16%
Otros	468	2%
Total	30,000	100%

Tabla 6. *Valores de **Educación**.*

Retomando lo comentado en la sección anterior tenemos que en la variable **Estado civil** existen valores que no vienen en la descripción de la base de datos, por lo tanto, se optó por agregarlos a “otros”, cabe mencionar que se tomó esta medida tomando en cuenta que el porcentaje que tiene “otros” con respecto del total es del 1% así como se muestra en la Tabla 7.

Estado civil	Frecuencia	Porcentaje
Casado	13,659	46%
Soltero	15,964	53%
Otros	377	1%
Total	30,000	100%

Tabla 7. *Valores de **Estado civil**.*

De la Tabla 8, se aprecia que la edad mínima para la aprobación de un crédito en el lapso en que se analizó fue de 21 años mientras que la máxima fue de 79 años, de la misma manera se analizó que la mayoría de las personas rondan los 35 años.

Edad	
Min	21
Max	79
Media	35.49

Desviación estándar	9.22

Tabla 8. *Valores de* ***Edad****.*

De la Figura 21 se reflejó que el punto máximo de frecuencia fue con una edad que ronda los 35 años y que el número de personas que tienen 60 años o más es el mínimo.

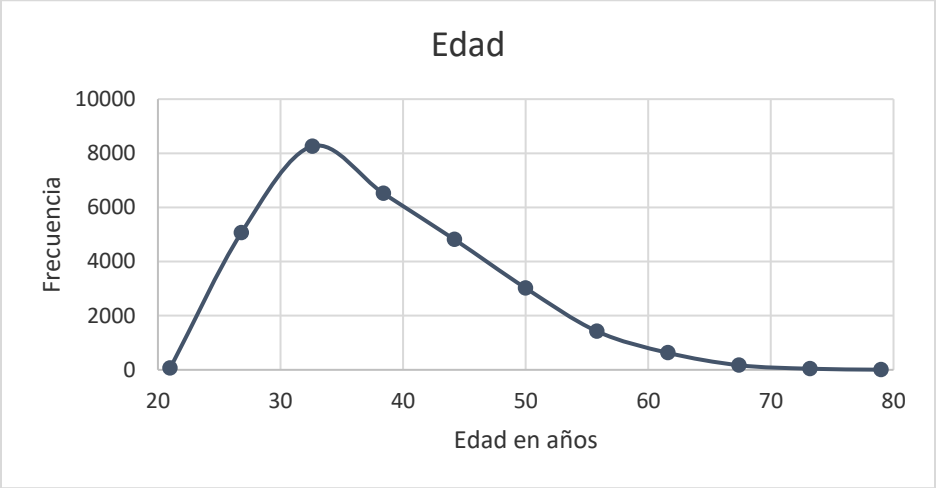


Figura 20. *Gráfica de la variable* ***Edad***.

A continuación, en la Figura 22 se presentan las gráficas de frecuencia de las variables **Pago i** con i = 1, 2, 3, 4, 5 y 6.

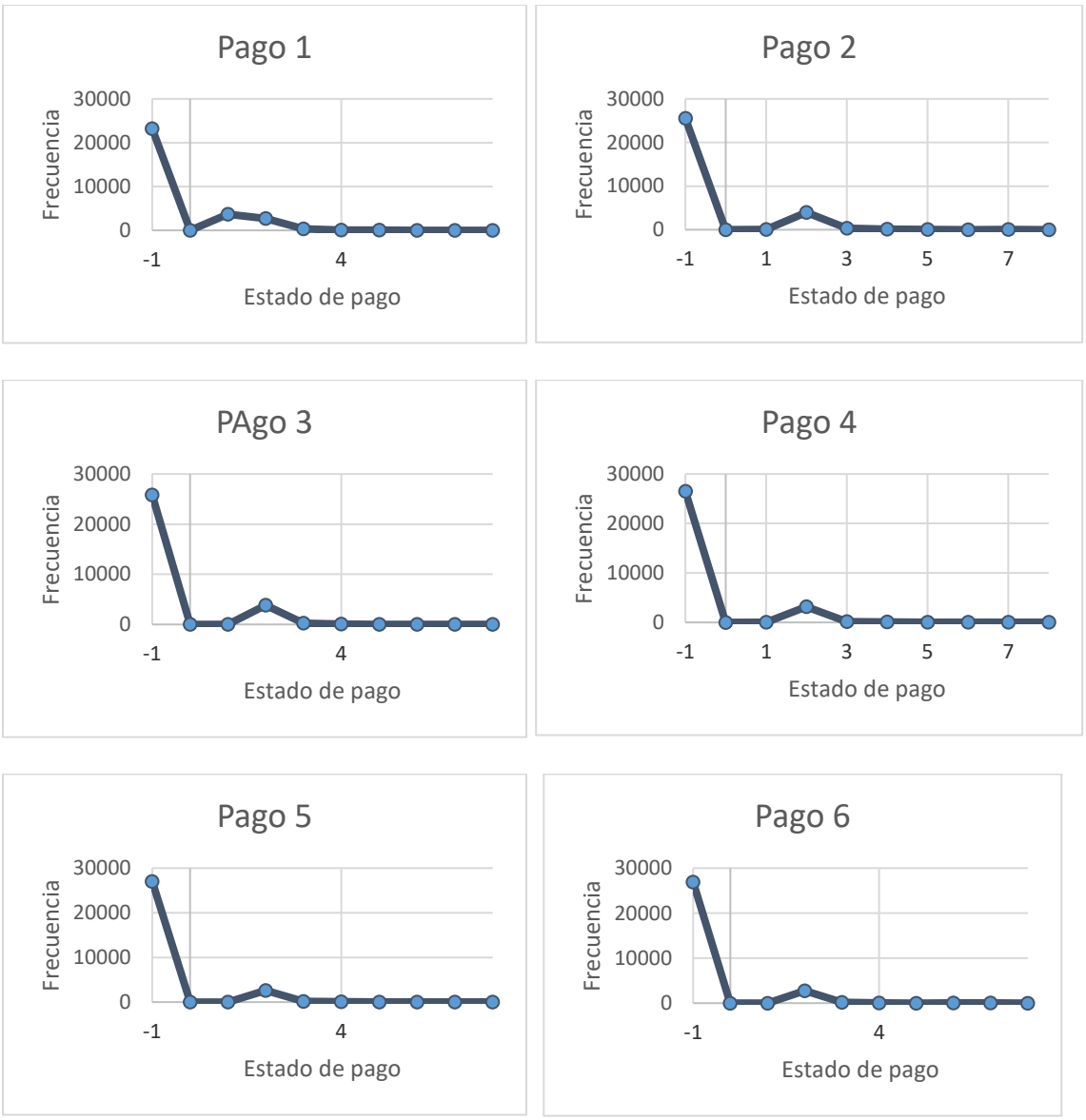


Figura 21. *Gráficas de frecuencia de las variables* ***Pago****.*

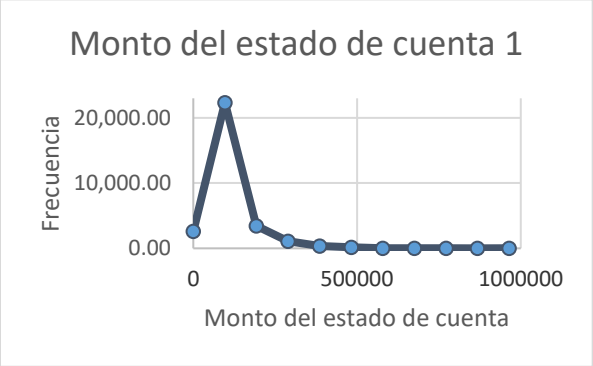
De la Figura 22 se encontró que a simple vista no existe una diferencia significativa de los pagos con respecto del tiempo. Para tener un mayor detalle sobre estas variables se presenta los valores en la Tabla 9.

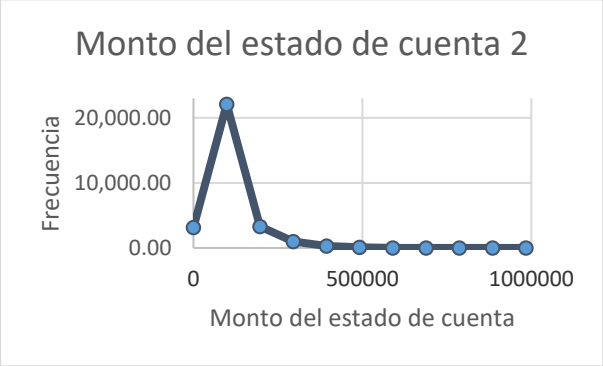
Variable	Pago 1	Pago 2	Pago 3	Pago 4	Pago 5	Pago 6
Mínimo	- 1.00	- 1.00	- 1.00	- 1.00	- 1.00	- 1.00
Máximo	8.00	8.00	8.00	8.00	8.00	8.00
Media	- 0.42	- 0.53	- 0.56	- 0.62	- 0.68	- 0.67
Desviación estándar	1.14	1.15	1.13	1.07	1.00	1.01

Tabla 9. *Valores de las variables* **Pago**.

De la Tabla 9 se analizó que no existe un cambio significativo entre mes y mes, pero se tuvo que en el mes 5 fue donde se hicieron más pagos debidamente hechos ya que tiene la media más cercana al -1.

En la Figura 23 se aprecia que en las variables **Monto del estado de cuenta i** con i = 1, 2, 3, 4, 5 y 6 tienen un comportamiento similar, es decir, tienen una distribución parecida.





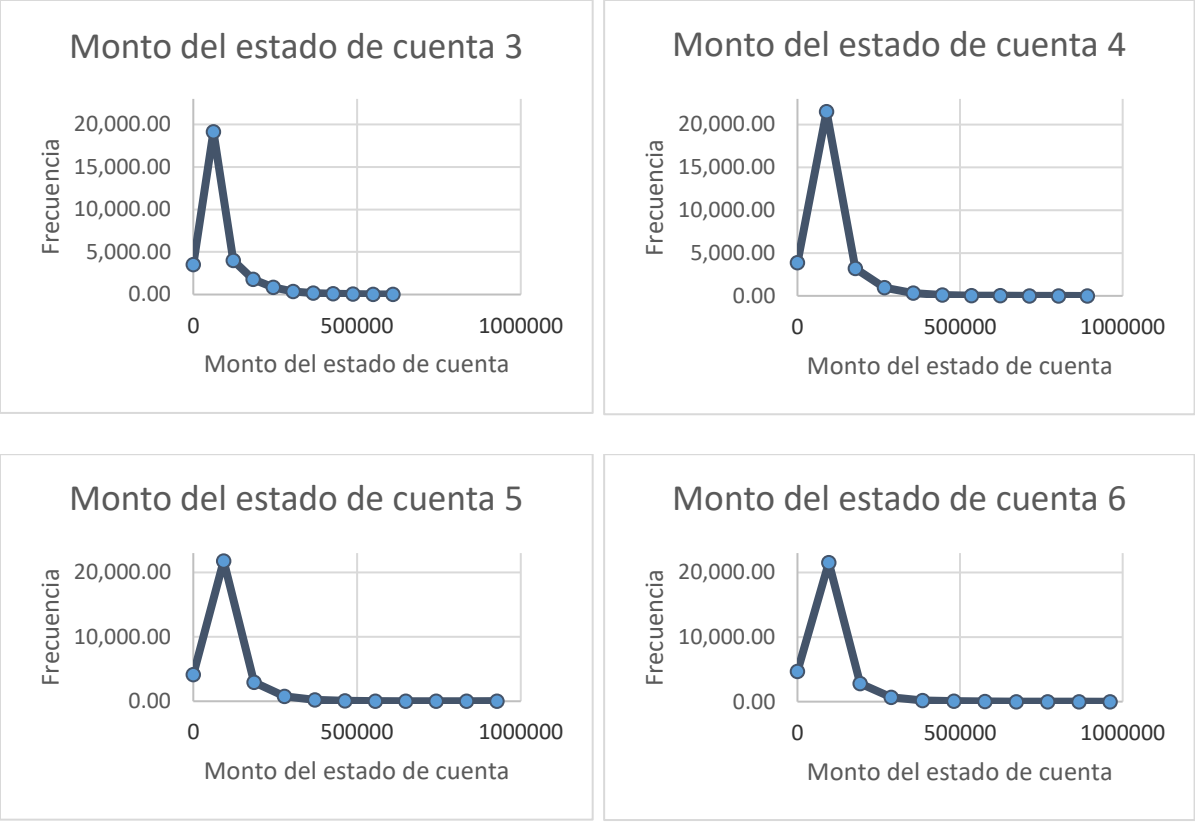


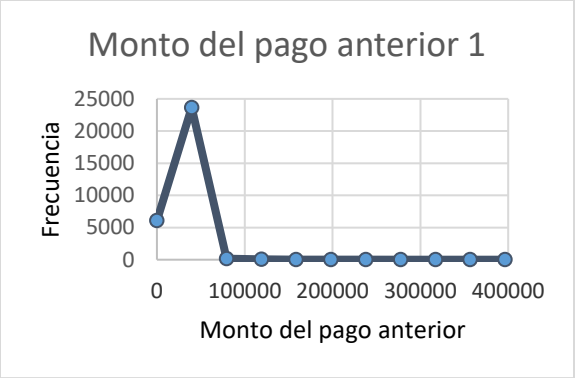
Figura 22. *Gráficas de frecuencia de las variables* ***Monto del estado de cuenta***.

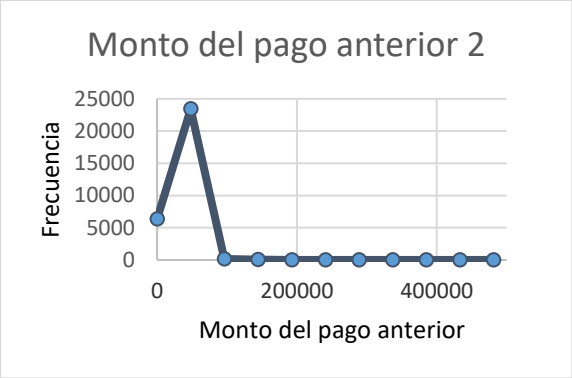
De la Tabla 10 analizamos que, de todos los meses, en donde hubo en promedio el estado de cuenta más bajo fue en el último, en el 6 y el más alto fue en el 1 que fue el que tuvo una mayor varianza.

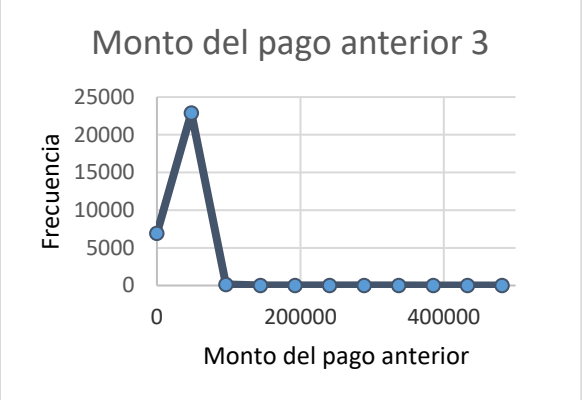
	Monto del estado de cuenta					
Variable	**1**	**2**	**3**	**4**	**5**	**6**
Mínimo	0	0	0	0	0	0
Máximo	964,511.00	983,931.00	610,000.00	891,586.00	927,171.00	961,664.00
Media	50,388.81	48,520.57	46,444.73	42,953.49	40,111.68	38,774.83
Desviación estándar	71,588.49	69,470.80	67,188.43	63,332.98	60,096.12	58,995.18

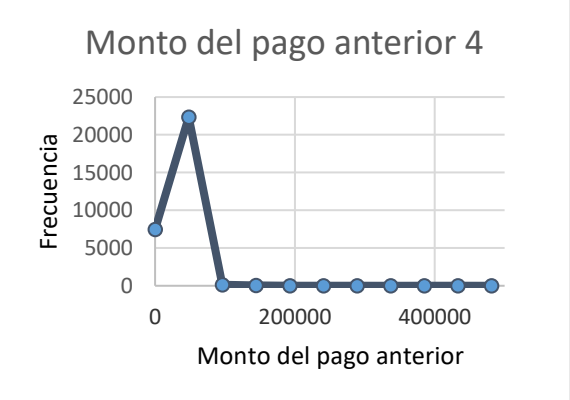
Tabla 10. Valores de las variables ***Monto del estado de cuenta***.

En la Figura 24, se presenta la gráfica de frecuencia de las variables **Monto del pago anterior i** con i = 1, 2, 3, 4, 5 y 6, al realizar el análisis de las gráficas, se tuvo que la distribución de estas variables son semejantes entre sí.









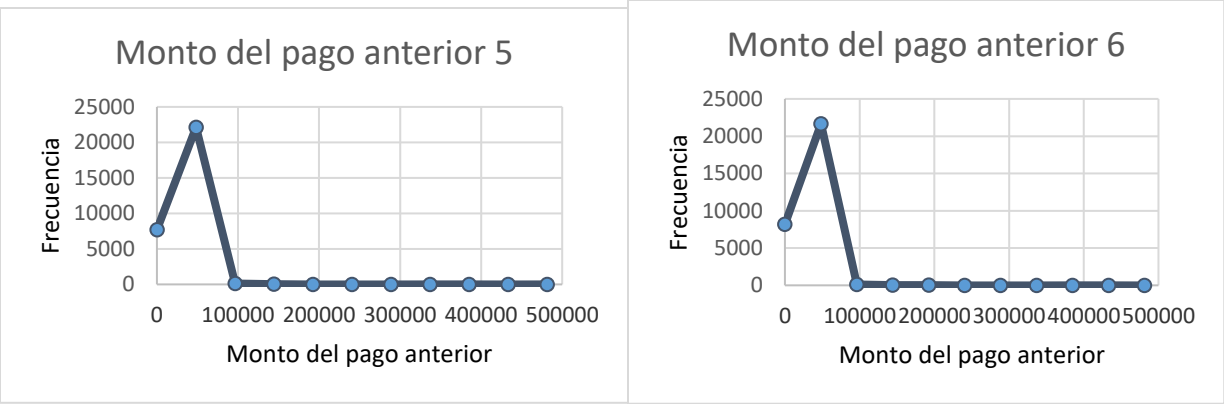


Figura 23. *Gráficas de frecuencia de las variables* ***Monto del pago anterior***.

Al analizar en conjunto las gráficas de **Monto del pago anterior** con la tabla de sus valores, se tiene que estas variables comparten características semejantes.

	Monto del pago anterior					
Variable	**1**	**2**	**3**	**4**	**5**	**6**
Mínimo	0	0	0	0	0	0
Máximo	396,343.00	481,382.00	535,020.00	432,130.00	400,993.00	443,001.00
Media	3,986.60	3,860.40	3,458.21	3,193.35	3,214.08	3,204.63
Desviación estándar	8,856.22	8,990.25	8,911.17	8,379.61	8,528.67	8,874.53

Tabla 11. *Valores de las variables* ***Monto del pago anterior***.

Por último, se analizó la variable dependiente, que es **Impago en el siguiente mes**, donde se encontró que el número de clientes morosos (6,636) es menor al de no morosos (23,364), esto de primera instancia es bueno para el banco ya que se esperaría que con ello la mayor parte del dinero haya sido pagada durante ese mes.

Impago en el siguiente mes	# clientes	clientes/total
# clientes de impago	6,636	22%
# clientes de pago	23,364	78%
Total	30,000	100%

Tabla 12. *Valores de* ***Impago en el siguiente mes****.*

Retomando el análisis anterior realizado sobre las variables se optó por construir nuevas a partir de las ya establecidas en la base de datos, esto se hizo con la finalidad de tener un mejor rendimiento en los clasificadores y también una mayor comprensión del significado de las variables respecto al peso que tiene con la variable dependiente.

La primera variable que se introdujo al modelo se le denominó **Mora**, la cual hace referencia al máximo retraso que tuvo el cliente en pagar la mensualidad, a continuación, se muestra la forma en que se calculó:

$$Mora = Máximo(Pago\ 1, Pago\ 2, Pago\ 3, Pago\ 4, Pago\ 5, Pago\ 6)$$

A continuación, en la Tabla 14 se presentan algunas características de la nueva variable, la cual era de esperarse que el mínimo de dicha variable sea el -1 haciendo referencia a que la persona siempre cumplió con sus pagos lo cual lo hace un cliente excelente y de la misma manera, existieron clientes morosos, los cuales dejaron de pagar durante 8 meses o más su crédito.

Mora	
Máximo	8.00
Mínimo	-1.00
Promedio	0.02
Desviación estándar	1.51

Tabla 13. *Valores de* ***Mora***.

La segunda variable se le denominó **%Uso** y como su nombre lo indica es el uso que tuvo el usuario de la tarjeta de crédito con respecto del límite máximo y se calculó de la siguiente forma:

$$\%Uso = \frac{\sum_{i=1}^{6} Monto\ del\ estado\ de\ cuenta_i}{Límite\ máximo}$$

Lo importante de esta variable es que se piensa que cuando una persona va a ser morosa se aprovecha de ello utilizando todo el dinero posible de su tarjeta de crédito, es decir utilizando toda la línea de crédito. A continuación, en la Tabla 14 se presentan algunos valores de la variable %Uso y se concluyó que la mayoría de las personas utilizan cerca del 37% del total de la línea de crédito.

%USO	
Máximo	1.00
Mínimo	0
Promedio	0.37
Desviación estándar	0.34

Tabla 14. *Valores %Uso*.

La última variable que se creó se llamó **%Pago** y expresa el porcentaje de dinero que pagó el usuario con respecto de su deuda, a continuación, se mostró la forma en que se calculó:

$$\%Pago = \frac{1}{6}\sum_{i=1}^{6}\left(\frac{Monto\ del\ pago\ anterior_i}{Monto\ del\ estado\ de\ cuenta_i}\right)$$

En la Tabla 15 se presentan algunos valores para la variable **%Pago** y se tiene que en promedio los clientes amortizan el 33% de su deuda mensualmente**.**

%Pago	
Máximo	1.00
Mínimo	-
Promedio	0.33
Desviación estándar	0.32

Tabla 15. *Valores %Pago*.

Teniendo en cuenta la información obtenida de forma individual de las variables independientes, se procedió a realizar gráficas que nos ayudaron a encontrar la relación que tienen las variables independientes con la dependiente.

A continuación, en la Figura 25 se presenta el gráfico de dispersión combinando las variables **Edad**, **Mora** y la variable dependiente **Impago en el siguiente mes**, con el propósito de encontrar la relación que existe entre ellas. A simple viste las variables **Edad** e **Impago en el siguiente mes** no parecen tener una relación muy fuerte, ya que, sin importar la edad del cliente, no se tiene claro si a mayor edad el cliente sea menos moroso o lo contrario, se concluye esto del gráfico porque el azul y el naranjado predominan en todas las edades mostradas, mientras que, por otro lado, tenemos la relación entre **Mora** e **Impago en el siguiente mes**, es claro que, cuando el cliente paga sus créditos en tiempo y forma se espera que pague su siguiente préstamo.

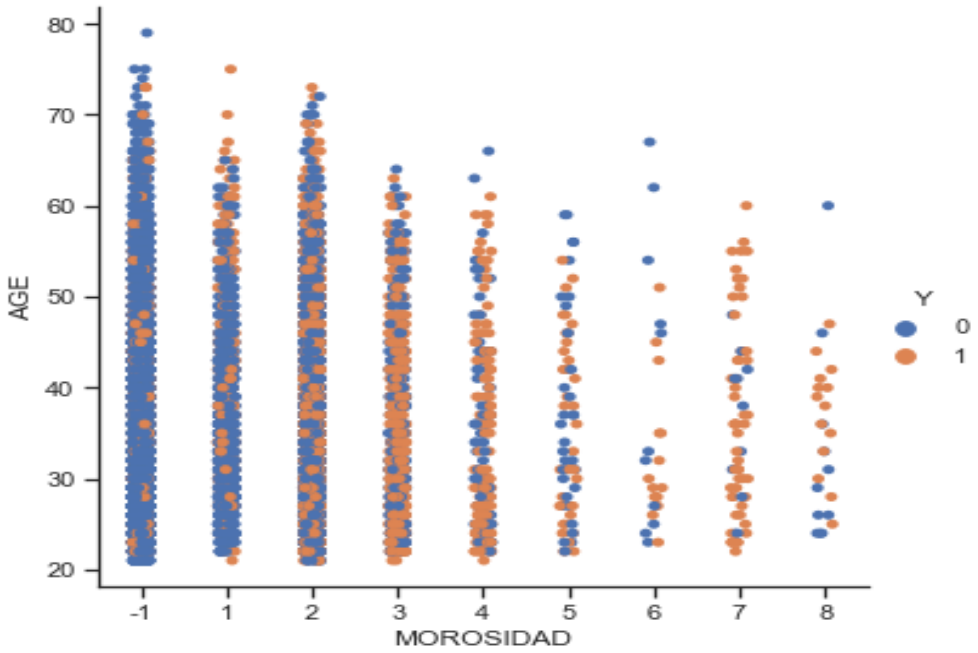


Figura 24. *Relación entre las variables **Edad** y **Mora***.

En la Figura 26 se presenta el gráfico de caja y brazos combinando las variables **Impago en el siguiente mes**, **Edad** y **Estado civil**, en este gráfico se puede ver como cambia la variable **Estado civil** respecto a la **Edad**, ya que el estado civil de una persona cuando está casada o se define como otros, la edad ronda entre los 40 y 50 años mientras que, cuando la persona está soltera ronda los 30.

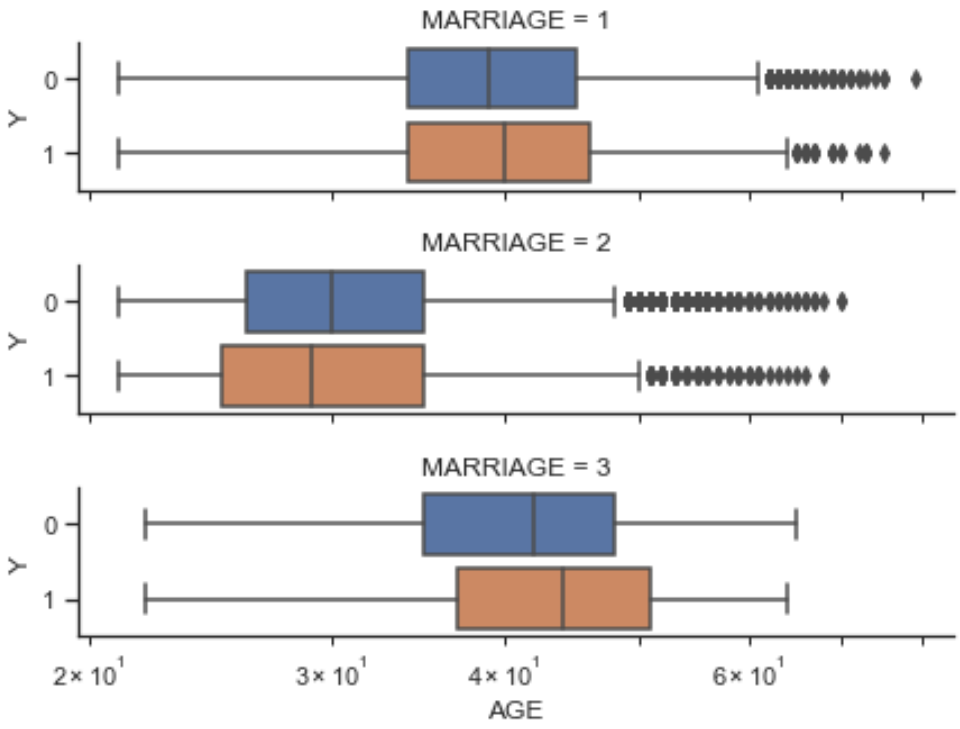


Figura 25. *Relación entre* ***Impago en el siguiente mes Edad y Estado civil.***

En la Figura 27, se presenta el gráfico de violín con las variables **Sexo**, **Límite máximo** e **Impago en el siguiente mes**, recordando que en **Sexo** = 1 hace referencia a hombre y **Sexo** = 2 a mujer. Lo relevante de este gráfico fue que en promedio los montos de préstamos más altos se otorgaron a mujeres y si fueron pagados.

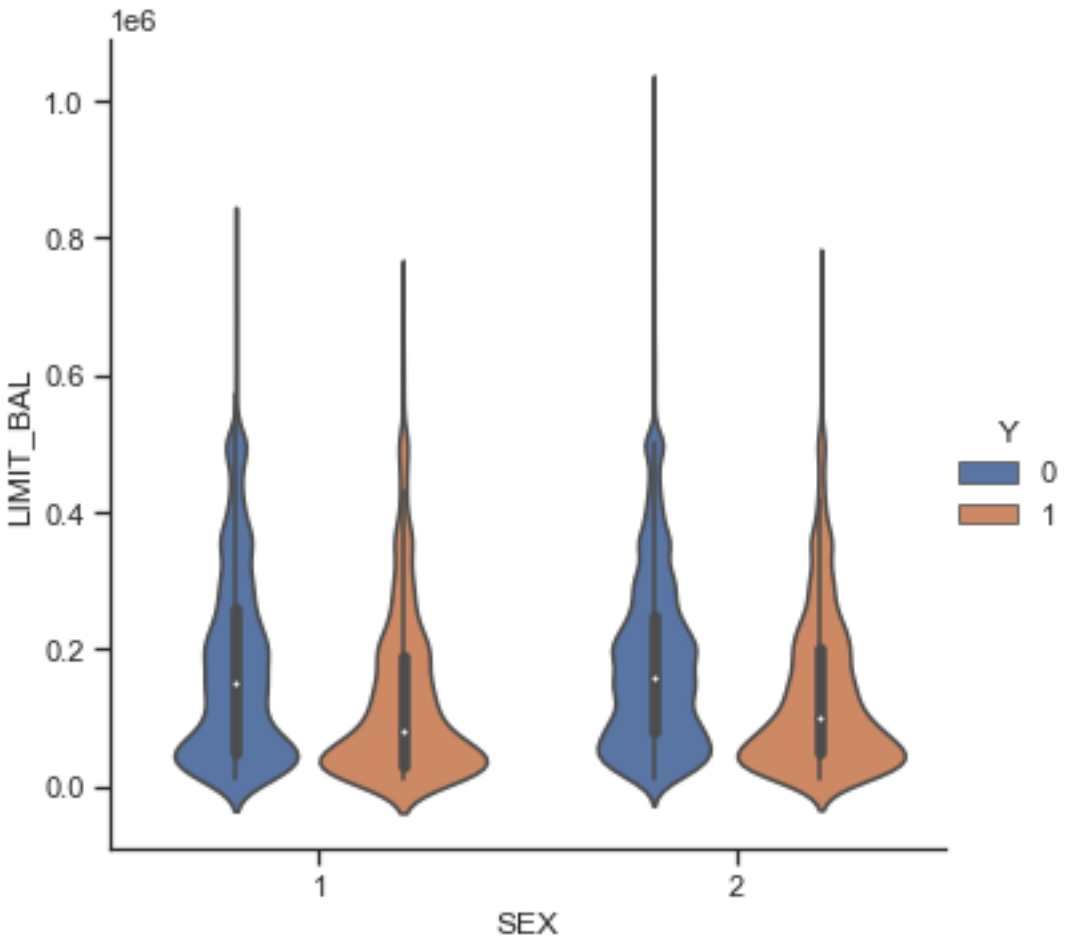


Figura 26. Gráfica de las variables Sexo, Límite máximo e Impago en el siguiente mes.

3.3 Construcción del clasificador

Ahora que se tiene un mayor conocimiento sobre las variables a evaluar podremos comenzar con la preparación de los datos y posteriormente a ello la elaboración de los clasificadores.

Los modelos que se aplicaron en este trabajo son los ya antes mencionados, modelo de regresión logística, análisis discriminante lineal, árbol de decisión y redes neuronales.

Para realizar los clasificadores se optó por utilizar el software libre Python, este lenguaje de programación es orientado al manejo de base de datos y es conocido por su amplia gama de herramientas que tiene para la ciencia de datos y un claro ejemplo está en la paquetería sklearn el cual contiene la mayoría de los modelos que se presentaron.

Se sabe que se puede dividir al conjunto de datos en dos, datos de entrenamiento y de prueba, para los clasificadores se utilizó el 80% de los datos para entrenamiento y 20% para prueba, cabe recalcar que la asignación de los datos de entrenamiento o de prueba se hace manera aleatoria con una semilla predeterminada, ya que en algunas ocasiones el orden de estos puede alterar la forma en que se construya el clasificador, a continuación, en la Tabla 16 se presenta la segmentación de los datos.

	# Observaciones	% Datos
Entrenamiento	24,000.00	80%
Prueba	6,000.00	20%
Total	30,000.00	100%

Tabla 16. Segmentación de la base de datos.

Para realizar esta segmentación se utilizó la librería *sklearn* con la semilla con el número 23, como a continuación se muestra:

$$X\_train, X\_test, y\_train\ , y\_test\ = train\_test\_split(X, y, test\_size = 0.2, random\_state = 23)$$

Se presenta en la Tabla 17 las variables que se utilizaron como independientes y la dependiente, dado que, en la sección 3.2 Análisis de las variables se crearon 3, las cuales se hicieron a partir de 19 variables que se encontraban en la base de datos, por lo tanto, de las 23 variables independientes se utilizaron solo 7.

Variables	
Independientes	**Dependiente**
Sexo	Impago en el siguiente mes
Educación	
Estado civil	
Edad	
Mora	
%Uso	
%Pago	

Tabla 17. Variables independientes y dependientes.

Para obtener mejores resultados en el clasificador se obtuvo el WoE por cada variable y como esta métrica solo se puede utilizar en variables categóricas se procedió a transformar las variables **Edad**, **%Uso** y **%Pago** para que de esta manera las 7 variables independientes del modelo fueran categóricas. Una vez obtenido el WoE por variable se sustituyeron todos los valores originales de la base de datos por los valores obtenidos por medio de esta métrica.

A continuación, en la Tabla 18 se presentan los intervalos que se utilizaron y los valores obtenidos por medio del WoE.

Educación	**WoE**
Posgrado	0.18
Universidad	-0.09
Secundaria	-0.17
Otros	1.32

Estado civil	**WoE**
Casado	-0.08
Soltero	0.07
Otros	-0.08

Sexo	**WoE**
Masculino	-0.12
Femenino	0.08

%Uso	**WoE**
[-0.001,0.0178)	-0.03
[0.0178,0.134)	0.53
[0.134,0.451)	0.28
[0.451,0.753)	-0.14

Edad	**WoE**
[20,25)	-0.25
[25,27)	0.08

Mora	**WoE**
En forma	0.76
Retraso 1	-0.16

[27,29)	0.14	Retraso 1	-1.00	[0.753,1)	-0.47
[29,31)	0.17	Retraso 1	-1.76		
[31,34)	0.17	Retraso 4	-1.84	**%Pago**	**WoE**
[34,37)	0.02	Retraso 5	-1.29	[-0.001,0.0427)	-0.34
[37,40)	0.07	Retraso 6	-1.50	[0.0427,0.0831)	-0.26
[40,43)	-0.04	Retraso 7	-2.89	[0.0831,0.357)	0.09
[43,49)	-0.08	Retraso +8	-1.50	[0.357,0.672)	0.38
[49,79)	-0.18			[0.672,1)	0.24

Tabla 18. Valores WoE.

3.4 Resultados

3.4.1 Regresión logística

En este modelo se tuvo en cuenta las 7 variables independientes anteriormente mencionadas, pero al elaborar diversas pruebas para dicho modelo se encontró que la variable **%Pago** no fue relevante, por lo tanto, se optó por omitirla.

Otro aspecto que sobresale de este modelo es que para realizar las predicciones se eligió un umbral de 0.4, esto con el propósito de mejorar su rendimiento discriminando, cabe mencionar que se llegó a este resultado después de haber realizado distintas pruebas, la cual se explica a lo largo de esta sección.

A continuación, se describieron las características más importantes de la Figura 27, la cual hace referencia a los resultados obtenidos por parte de este clasificador:

- **Dep. Variable**: Es la variable dependiente, es decir **Impago en el siguiente mes**, la cual se abrevió como “Y”.
- **Method**: Es el método empleado para obtener los coeficientes del modelo, “MLE” hace referencia Máxima Verosimilitud.

- **No**. **Observations**: Es el número de observaciones empleadas para crear el modelo, los cuales hacen referencia a las 24,000 muestras es decir los datos de entrenamiento.
- **DF Residuals**: Es el número de observaciones (24,000) menos el número de parámetros que se están estimando (7).
- **DF Model**: son los grados de libertad, hace referencia al número de variables independientes que se usan en el modelo, esto con la intención de hacer pruebas estadísticas.
- **Pseudo R-squared**: Hace referencia a la bondad de ajuste del modelo a los datos.

Logit Regression Results

Dep. Variable:	Y	No. Observations:	24000
Model:	Logit	Df Residuals:	23993
Method:	MLE	Df Model:	6
Date:	Thu, 30 Sep 2021	Pseudo R-squ.:	0.1369
Time:	10:42:57	Log-Likelihood:	-10983.
converged:	True	LL-Null:	-12725.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2462	0.017	-72.429	0.000	-1.280	-1.213
SEX	-0.6157	0.176	-3.497	0.000	-0.961	-0.271
EDUCATION	-0.4666	0.093	-4.995	0.000	-0.650	-0.283
MARRIAGE	-1.1558	0.233	-4.954	0.000	-1.613	-0.699
AGE	-0.3964	0.124	-3.196	0.001	-0.639	-0.153
MOROSIDAD	-0.9609	0.019	-51.396	0.000	-0.998	-0.924
%USO	-0.5084	0.050	-10.118	0.000	-0.607	-0.410

Figura 27. *Anova de Reg logit*

En la parte inferior de la Figura 28 se tiene una tabla, donde la primera columna hace referencia a las variables independientes y los títulos hacen referencia a los nombres de los estadísticos aplicados al modelo, los cuales se tiene que:

- **Coef (coeficiente)** es el valor por el que se multiplica la muestra.
- **Std err (desviación estándar)** es la estimación de la varianza del coeficiente

- **z (estadístico z)**, es un estadístico que mide la diferencia que existe entre su valor hipotético y su parámetro observado.
- **P>|z| (P-Value)**, es una prueba estadística que sirve para ver la significancia de la variable en el modelo, generalmente se usa al 5%
- **[0.025, 0.975] (Intervalo de confianza)**, es un intervalo en el que se espera que se encuentre el valor real del coeficiente, el intervalo mostrado anteriormente utiliza un nivel de significancia de 5%.

Por medio del estadístico P-Value fue fácil analizar que la variable **%Pago** no tuvo relevancia en el modelo, ya que, obtuvo un P-Value de 0.149 el cual es mayor al 5% estipulado anteriormente.

Tomando en cuenta la información obtenida por parte de la Figura 28, se obtuvo que la variable **Estado civil** tuvo el mayor error estándar, esto se puede ver reflejado en el intervalo de confianza ya que este es el intervalo más grande de todas las variables.

Dado que se introdujeron valores positivos y negativos en cada variable al modelo (Tabla 18), no es tan sencillo analizar los criterios que toma como bueno o malo, un ejemplo está en la variable independiente con mayor peso que es **Estado civil**, se tiene que el único valor positivo en dicha variable es “soltero”, dado que, el coeficiente es negativo se tiene que los solteros tienden a ser más morosos, por el otro lado tenemos que para los “casados” tienen a pagar más sus créditos.

A continuación, en la tabla 19 se presentan las predicciones de este modelo con respecto de todo el conjunto de datos, es decir fueron 30,000 predicciones de las cuales se clasificaron por los valores reales los clientes buenos y malos, esto con el objetivo final de realizar diversas pruebas.

Score Agrupado	# Clientes buenos	# Clientes malos	%Buenos	%Malos	WoE	IV
0.00	2712.00	2786.00	11.61%	41.98%	-1.29	0.39
107.25	1408.00	922.00	6.03%	13.89%	-0.84	0.07
214.49	461.00	221.00	1.97%	3.33%	-0.52	0.01
321.74	697.00	258.00	2.98%	3.89%	-0.26	0.00
428.99	647.00	154.00	2.77%	2.32%	0.18	0.00
536.24	3529.00	638.00	15.10%	9.61%	0.45	0.02
643.48	7764.00	1097.00	33.23%	16.53%	0.70	0.12
750.73	5132.00	474.00	21.97%	7.14%	1.12	0.17
857.98	907.00	84.00	3.88%	1.27%	1.12	0.03
1,072.47	107.00	2.00	0.46%	0.03%	2.72	0.01
Total	**23,364.00**	**6,636.00**	**100%**	**100%**	**IV**	**0.82**

Tabla 19. Information Value Logit.

De la Tabla 19 se tiene que el score obtenido de cada predicción se agrupó en 11 intervalos que va desde el 0 hasta el 1,072.47, tomando en cuenta que la diferencia entre todos ellos es de 107.25, donde todas las predicciones negativas e iguales a 0 se agrupó en el máximo de todos estos, el cual es el 0, todos los valores mayores a 0 y menores o iguales a 107.25 agruparon en 107.25 y así sucesivamente, un detalle a resaltar es que en el último grupo que es de 1,072.47 se unieron los dos últimos intervalos, esto con la intención de que en todos los intervalos existan clientes buenos y malos.

A continuación, se presentan las fórmulas utilizadas para realizar la Tabla anterior:

- **%Buenos** es la razón que tiene cada número de clientes buenos de cada score con respecto del total de clientes buenos, es decir: $\%Buenos = \frac{\%Buenos_i}{Total\ de\ \%Buenos}$
- **%Malos** es la razón que tiene cada número de clientes malos de cada score con respecto del total de clientes malos, es decir: $\%Malos = \frac{\%Malos_i}{Total\ de\ \%Malos}$
- **WoE** que significa Weight of Evidence del español peso de la evidencia se utilizó la siguiente formula: $WOE = ln\left[\frac{Distribución\ de\ buenos}{Distribución\ de\ malos}\right]$

- **IV** se obtuvo como la diferencia de %AcumBuenos y %AcumMAlos multiplicado por el WoE, es decir: $IV = \sum(\%\ acumulado\ de\ buenos\ - \%\ acumulados\ de\ malos\ ) \cdot WOE$

Para obtener la métrica *Information Value* se sumó toda la columna IV de la Tabla 19, la cual dio un resultado de 0.82, con este resultado no se puede llegar a tener una conclusión sobre esta métrica ya que se encuentra en el rango de lo dudoso y se tuvo que en el score de 0 se obtuvo un 0.39 debido a la gran cantidad de porcentaje que se tiene en el acumulado de los clientes malos con respecto de su total en comparación con el acumulado de los clientes buenos con su total.

Score Agrupado	%Acum Buenos	%Acum Malos	Delta Buenos	Delta Malos	Ai
-	11.61%	41.98%	0	0	-
107.25	17.63%	55.88%	29.24%	13.89%	0.020
214.49	19.61%	59.21%	37.24%	3.33%	0.006
321.74	22.59%	63.10%	42.20%	3.89%	0.008
428.99	25.36%	65.42%	47.95%	2.32%	0.006
536.24	40.46%	75.03%	65.82%	9.61%	0.032
643.48	73.69%	91.56%	114.16%	16.53%	0.094
750.73	95.66%	98.70%	169.35%	7.14%	0.060
857.98	99.54%	99.97%	195.20%	1.27%	0.012
965.23	100.00%	100.00%	199.54%	0.03%	0.000
				Total	**0.239**

Tabla 20. Índice de Gini Logit.

A continuación, se presentan las fórmulas utilizadas para obtener los valores de la Tabla 20:

- **Delta Buenos** es la suma de %Acum Buenos más la posición anterior, es decir**:** $Delta\ Buenos\ = \%Acum\ Buenos_i + \%Acum\ Buenos_{i-1}$
- **Delta Malos** es la diferencia entre %Acum Malos con la posición anterior, es decir: $Delta\ Malos\ = \%Acum\ Malos_i - \%Acum\ Malos_{i-1}$

- **Ai** es el producto de DeltaBuenos con DeltaMalos y 0.5, es decir: $Ai = DeltaBuenos_i * DeltaMalos_i * 0.5$

Para crear la métrica índice de Gini se utilizó la siguiente formula:

$$IG = \frac{AT - AG}{AT} = \frac{0.5 - 0.239}{0.5} = 0.521$$

Donde AT es igual a 0.5 que hace referencia a la mitad del área de un cuadrado de área 1 y AG es igual a la suma de la columna Ai. Para esta métrica se analizó que el score agrupado 643.48 fue el que mayor aportó valor, esto debido al gran incremento que se tuvo en Delta Buenos en esa sección.

Score Agrupado	%Acum Buenos	%Acum Malos	Delta (B-M)
-	11.61%	41.98%	0.30376
107.25	17.63%	55.88%	0.38243
214.49	19.61%	59.21%	0.39600
321.74	**22.59%**	**63.10%**	**0.40505**
428.99	25.36%	65.42%	0.40056
536.24	40.46%	75.03%	0.34566
643.48	73.69%	91.56%	0.17867
750.73	95.66%	98.70%	0.03044
857.98	99.54%	99.97%	0.00428
965.23	100.00%	100.00%	0.00000

Tabla 21. Kolmogorov-Smirnov Logit.

De la Tabla 21 se tiene que el punto de corte del score se encuentra en 321.74, esto debido a que obtuvo el mayor delta con un 0.4050, es decir, en este intervalo se encontró la mayor discrepancia entre el acumulado de buenos clientes con el acumulado de malos clientes, por lo tanto, el umbral de 0.4050 fue el que se utilizó para poder realizar las clasificaciones.

		Modelo	
		0	**1**
Real	**0**	4,062	645
	1	693	600

Tabla 22. Matriz de confusión regresión logit.

Retomando lo comentado en la sección 2.4 tenemos que en la diagonal de la matriz (los valores sombreados en verde) de la Tabla 22 se tienen los valores acertados de la predicción, es decir, el primer elemento de este es 4,062 y es el número de predicciones que realizó el modelo a los clientes no morosos y que fueron correctas, lo mismo pasa para el segundo elemento de la identidad, 600 es el número de morosos que predijo el modelo y que en verdad lo fueron.

La posición (1,2) donde se encuentra el 645 hace referencia al número total de predicciones a los clientes como morosos, pero en realidad no lo fueron, es decir al error tipo1 y en la posición (2,1) con el 693 es el número de predicciones a los clientes como no morosos, pero en realidad si lo fueron, o sea al error tipo 2.

De la Tabla 23 se tiene que el número de aciertos totales que tuvo el modelo fue de 4,662 y de errores 1,338 lo cual de primera instancia se tiene que es aceptable por la gran diferencia que las predicciones correctas tienen sobre los errores, pero para tener una conclusión solida sobre el modelo se presenta en la Tabla 23 algunas métricas de este modelo, cabe recalcar que 3 de las 4 presentadas a continuación se obtienen a partir de la matriz de confusión.

ROC-AUC	Exactitud	Especificidad	Sensibilidad
66.35%	77.70%	86.30%	46.40%

Tabla 23. Métricas regresión logit.

En la Tabla 23 se presentaron las principales métricas en el ámbito de la inteligencia artificial y como se analizó en la sección 2.4.2 tenemos que:

- ROC-AUC hace referencia a la capacidad de discriminación que tiene un modelo con respecto de dos poblaciones siendo el 1 la máxima capacidad y 0.5 la mínima.
- Exactitud es el número de predicciones correctas entre el número total de observaciones, es decir $\frac{4{,}062 + 600}{4{,}062 + 600 + 693 + 645}$
- Especificidad es el número de predicciones correctas de los clientes no morosos entre el número total de los clientes reales no morosos, es decir $\frac{4{,}062}{4{,}062 + 645}$
- Sensibilidad es el número de predicciones correctas que se hizo de los clientes morosos entre el total de los morosos reales, es decir $\frac{600}{600 + 693}$

Partiendo de las explicaciones anteriormente dadas, es fácil inferir que, a mayores valores arrojados por cada métrica hace referencia a mejores rendimientos, por lo tanto, al analizar las métricas se encontró que la prueba ROC-AUC estuvo cerca de lo aceptable porque no alcanzó el 70%, pero por otro lado se obtuvo por medio de la exactitud un valor de 77.70%, lo cual es bastante bueno.

Las otras métricas a analizar son la especificidad y la sensibilidad del modelo, estas métricas son la razón que tienen los valores acertados del modelo entre los valores totales reales por cada categoría, es decir se especificidad se obtuvo un 86.30% y da a entender que se acertó en la mayoría de las predicciones para los clientes no morosos, mientras que para la sensibilidad se tuvo el 46.40%, lo cual es un porcentaje muy bajo y hace referencia que para los clientes morosos el clasificador no tuvo un buen desempeño.

3.4.2 Redes neuronales

Para la elaboración de este modelo se realizaron diversas pruebas para obtener los mejores resultados, ya que, existen varios factores al momento de realizar dicho clasificador, por ejemplo, se debe de elegir las funciones de activación para cada capa neuronal y aunado a ello, se debe de tomar en cuenta el número de neuronas que se desean.

La construcción de este modelo es relativamente simple, dado que, de entrada, se tiene las 7 variables independientes que se conectan con una capa oculta que tiene 15 neuronas con la función de activación relu y, por último, en la capa de salida se tiene una neurona con la función de activación *sigmoid*, el resultado de esta función serán valores entre el 0 y el 1 y para saber si el modelo clasificó al cliente como bueno o malo se propuso un umbral con un valor de 0.5, así de esta manera si el valor arroja es menor al umbral se tomará como cliente bueno, de lo contrario como moroso.

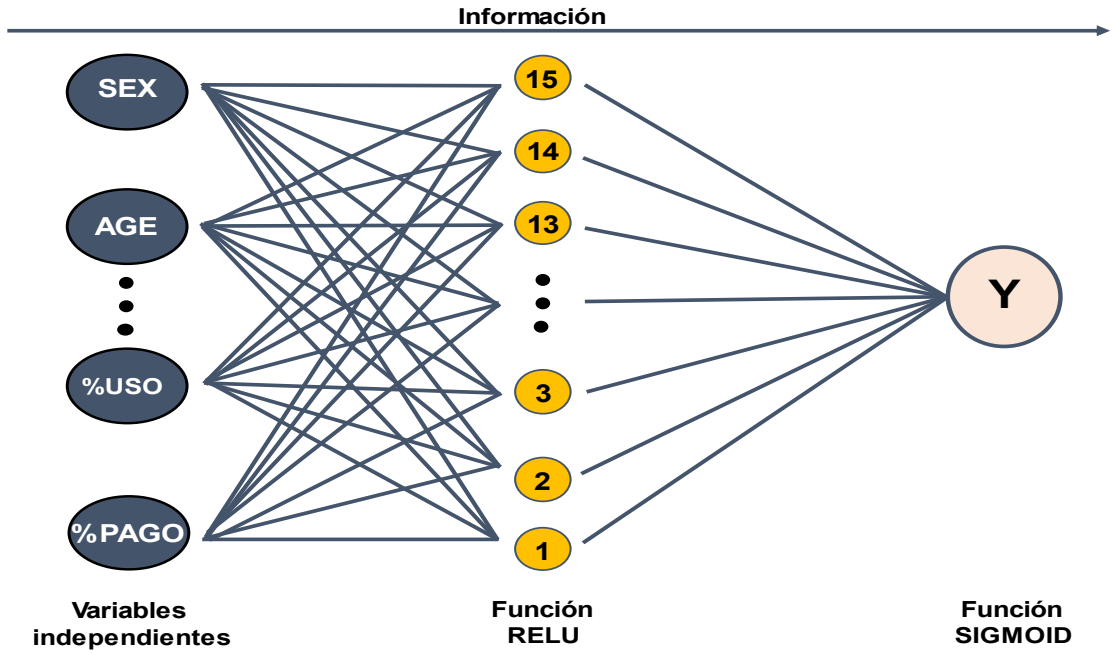


Figura 28. Redes Neuronales.

En la Figura 28 se mostró el esquema de lo comentado, la primera columna de óvalos azules es la capa de entrada con las variables **Sexo**, **Educación**, **Estado civil**, **Edad**, **Mora**, **%Uso** y **%Pago**, la segunda capa es la única oculta que está representada por los 15 óvalos amarillos y al final la capa de salida.

Se presenta en la Tabla 24 la matriz de confusión de este modelo aplicado a los datos de prueba.

		Modelo	
		0	**1**
Real	**0**	2,249	113
	1	498	140

Tabla 24. *Matriz de confusión redes neuronales*.

De la matriz de confusión se tiene que el número de predicciones correctas del modelo fue de 2,249 y 140 para los clientes buenos y los malos respectivamente, es decir, el modelo tuvo un total de 2,389 de predicciones correctas y un total de 611 de errores. Teniendo en cuenta los valores arrojados de la matriz de confusión se presenta a continuación algunas métricas, las cuales 3 de las 4 presentadas se obtienen a partir de dicha matriz.

ROC-AUC	Exactitud	Especificidad	Sensibilidad
58.57%	79.63%	95.22%	21.94%

Tabla 25. *Métricas redes neuronales.*

De la Tabla 25 se obtuvo que el rendimiento en la prueba ROC-AUC no fue aceptable, ya que dicho valor se encuentra cercano al 50% y un reflejo este resultado se encuentra en las últimas dos métricas presentadas, porque por un lado tenemos que la especificidad fue de 95.22% pero la sensibilidad del modelo fue de 21.94% lo cual dice que para predecir a los clientes morosos no se tuvo una buena discriminación, pero por otro lado en la métrica de exactitud fue buena y hace referencia que de todas las predicciones del clasificador el 79.63% fueron acertadas.

3.4.3 Árbol de decisión

Este clasificador tiene como criterio la ganancia de informacion (entropy), el cual tiene como función segmentar los nodos, este proceso por lo general se hace hasta tener la profundidad máxima del árbol y en este trabajo se espera que tenga una profundidad relativamente grande porque son 7 variables independientes con 24,000 muestras, pero en este caso para una mejor comprensión y visualización del modelo se optó por tener una profundidad de 3.

En la Figura 29 se presenta de manera esquemática la forma en que funciona este modelo, se puede apreciar cuales son los valores entropy para segmentar el árbol y como un punto importante a observar del árbol, es que no es una regresión, es un clasificador y por lo tanto en la última condición que este presenta, son los resultados del clasificador, es decir, los únicos resultados que puede dar es que si el cliente es bueno o malo. Cabe resaltar que en el modelo se pueden apreciar que 6 de los 8 ultimas condiciones hacen referencia a que el cliente es bueno, esto debido al desbalanceo que existe en la base de datos, pero esto no es malo ya que al llevarlo a la práctica se espera que la mayoría de los clientes sean buenos y esto por el proceso de filtros que se les aplican para otorgarles la tarjeta de crédito.

A continuación, se presentan algunas características de este árbol que se encuentran en la Figura 29:

- La condición que se encuentra arriba de cada rectángulo es la regla que se ocupa para saber cuál es la siguiente condición que le toca a la muestra según sus características.
- **Entropy** hace referencia a la función que se ocupa para poder elegir la mejor condición.
- **Samples** es el número de muestras que cumplen con las condiciones de cada nodo.
- **Value** son los valores en que se dividen las muestras para probar la ganancia de información.
- **Class** hace referencia a la clase en la que se encuentra la muestra, ya sea bueno, es decir que el cliente si paga o malo, que es moroso.

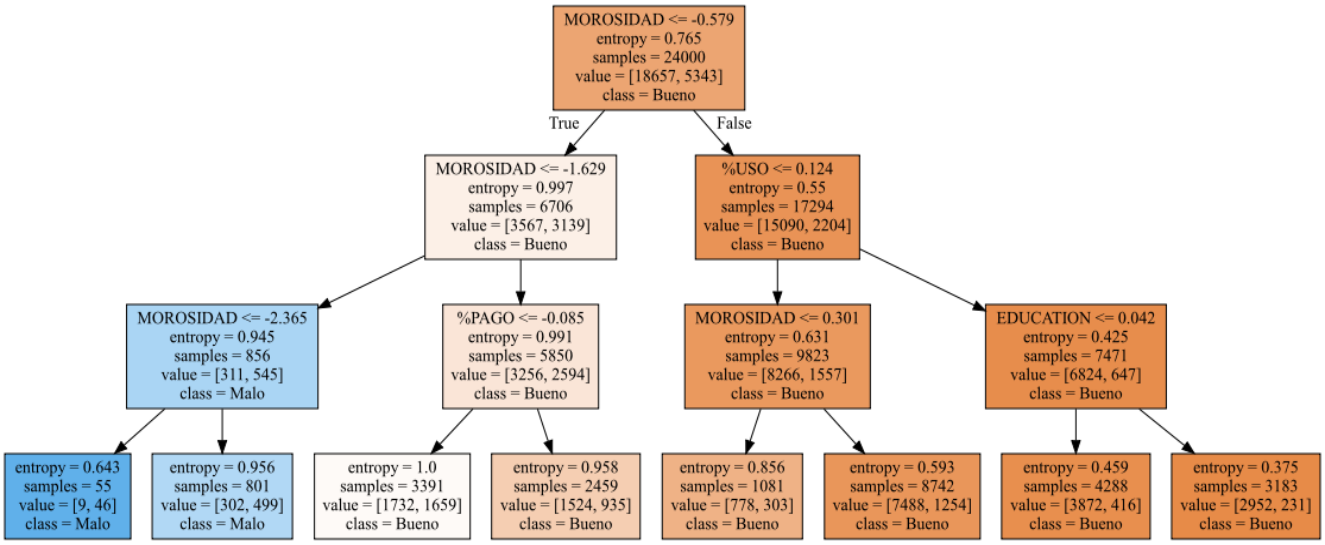


Figura 29. Esquema árbol de decisión.

En la figura 29 se elaboró con colores para poder analizar con más facilidad las categorías, ya que, el azul hace referencia a un cliente malo y el naranja al bueno. En la Tabla 26 se presenta la matriz de confusión de este modelo aplicados a los datos de entrenamiento.

		Predicción	
		0	**1**
Real	**0**	4,631	76
	1	1,151	142

Tabla 26. Matriz de confusión árbol de decisión.

De la Tabla 26 se tiene que el modelo tuvo un total de 4,773 de predicciones correctas y 1,247 de predicciones erróneas, de primera instancia se tiene que es relativamente bueno el modelo ya que supera en gran cantidad el número de buenas observaciones que al de los errores, pero para decidir si es un buen modelo se tomaron en cuenta otras métricas. A continuación, se presentaron métricas de este modelo y algunas fueron obtenidas a partir de la matriz de confusión.

ROC-AUC	Exactitud	Especificidad	Sensibilidad
54.68%	79.55%	98.39%	10.98%

Tabla 27. *Métricas árbol de decisión*.

De la Tabla 27 se analizó el desempeño que tuvo este modelo y se observó que en ROC-AUC se tuvo un rendimiento muy bajo junto con el de sensibilidad, pero por otro lado en la exactitud y especificidad del modelo se obtuvieron buenos resultados, cabe mencionar que el bajo desempeño en ROC-AUC se puede ver reflejado por la baja capacidad de discriminación en la métrica sensibilidad y de la misma manera si el modelo obtuvo un rendimiento aceptable en exactitud fue por la buena capacidad de discriminación en especificidad.

3.4.4 Análisis discriminante lineal

En la Tabla 28 se presenta la matriz de confusión de este modelo aplicado a los datos de prueba, donde los valores subrayados en verde son las predicciones correctas de dicho modelo y las que no están sombreadas son los errores obtenidos.

		Modelo	
		0	**1**
Real	**0**	4,243	464
	1	818	475

Tabla 28. Matriz de confusión análisis discriminante lineal.

Se tiene que el modelo tuvo un total de aciertos de 4,718 donde 4,243 son para los clientes no morosos y 475 para los morosos y un total de errores de 1282 donde 818 hacen referencia al error tipo 2 y 464 al error tipo 1. En la siguiente tabla se presentaron algunas métricas aplicadas a este modelo, las cuales tres de ellas se obtienen a partir de la matriz de confusión.

ROC-AUC	Exactitud	Especificidad	Sensibilidad
63.43%	78.63%	90.14%	36.73%

Tabla 29. Métricas redes neuronales.

Al analizar la Tabla 29 se encontró que, la métrica ROC-AUC no tuvo un desempeño tan bajo como en los modelos anteriores, pero sin duda alguna no sigue siendo aceptable, de la misma manera se puede ver reflejó con la sensibilidad del clasificador ya que, en este caso fue bajo pero no tanto a comparación de los modelos anteriores, por parte de las métricas exactitud y especificidad se tuvieron buenos resultados, cabe mencionar que de nuevo la métrica exactitud sigue teniendo un buen resultado por parte de la especificidad, el cual hace referencia a la buena predicción sobre los clientes no morosos.

3.4.5 Comparación entre los modelos

Se presenta a continuación en la Tabla 31 un resumen de las métricas mostradas por cada clasificador con el fin de poder compararlos directamente, para mayor facilidad del análisis sobre la tabla se subrayó en verde los máximos valores arrojados por cada métrica y en naranja los mínimos valores.

Métricas\Modelo	ROC-AUC	Exactitud	Especificidad	Sensibilidad
Logit	66.35%	77.70%	86.30%	**46.40%**
Redes	58.57%	**79.63%**	95.22%	21.94%
Árbol	54.68%	79.55%	**98.39%**	10.98%
ADL	**63.43%**	78.63%	90.14%	36.73%

Tabla 30. Tabla resumen.

De la Tabla 30 se tiene que ninguno de los modelos presentados sobresale de los demás, es decir, no existe alguno que en todas las métricas sea mejor que los otros, pero se tuvieron valores muy interesantes ya que el modelo de análisis discriminante tuvo el mejor rendimiento en ROC-AUC, redes neuronales fue el mejor en la exactitud, el árbol de decisión en la especificidad y por último el de regresión logística fue el mejor en la sensibilidad.

Quizá las cuatro métricas presentadas de los modelos no tienen el mismo peso, pero incluso así obtuvieron valores muy cercanos entre sí, es decir si tomamos la máxima diferencia que existe entre los cuatro valores que se tiene en exactitud es de tan solo 1.93, lo cual no resulta ser significativo como obtener una conclusión de aquí, por lo tanto, se optó por tener una perspectiva completa sobre cada modelo para elegir al mejor.

Para elegir el mejor modelo para esta base de datos se analizó la dificultad que se tuvo para la comprensión , elaboración y presentación y debido a ello se optó por el clasificador de regresión logística, el cual para poder comprenderlo solo se necesita conocimientos en estadística básica, para su elaboración de igual manera resulta muy práctico porque no se ocupan parámetros que compliquen su elaboración y la razón más importante se encuentra en su presentación, la cual se puede modificar para que cualquier persona pueda utilizar el modelo sin necesidad de tener bases en estadística, lo cual esto en el ámbito laboral es de gran utilidad.

Conclusiones

Al inicio de la tesis se planteó encontrar el mejor modelo de entre los presentados en este trabajo para clasificar a los clientes que usan la tarjeta de crédito de un banco entre buenos y malos y por medio de las métricas presentadas se observó que el rendimiento de cada uno de los modelos fueron aceptables, sin embargo, los valores arrojados son muy cercanos entre cada uno de los modelos, por lo que se tornó más complicado para poder escoger el mejor clasificador y se optó por tener un panorama más amplio, es decir, no solo se tomó en cuenta los resultados de cada clasificador, si no, también el nivel de dificultad que tiene cada uno y la forma en que estos se pueden presentar.

Otro aspecto importante para profundizar es en el nivel de aversión al riesgo en que se encuentre la institución, esto con el propósito de ajustar el modelo a lo deseado y un claro ejemplo está en proponer el umbral del clasificador, si se tiene que 0 es no moroso y 1 moroso y la institución es muy adversa al riesgo se recomienda tener el umbral arriba del 0.5 y por el otro lado si tiene un apetito al riesgo elevado entonces el umbral debería de estar abajo del 0.5.

Los cuatro modelos presentados arrojaron rendimientos aceptables, lo cual resultó ser muy bueno porque la base de datos utilizada no fue la más exacta para la elaboración de un clasificador y esto se puede ver reflejado en las predicciones enfocadas a los clientes morosos que resultó ser de bajo rendimiento, pero si se hubiera tenido mayor y mejor información sobre los clientes como el tipo de empleo, salario o número de hijos se esperaría que los resultados de los clasificadores fueran idóneos para llevarlos a la práctica.

Después de analizar lo comentado anteriormente se optó por el modelo de regresión logística por dos grandes razones, la primera hace referencia a que solo se necesita estadística básica para su fácil comprensión y manejo del mismo, la segunda se encuentra en la forma de presentarlo, ya que como se mostró anteriormente se puede realizar mediante un resumen general del modelo el cual es el anova o utilizando métricas enfocadas a los clasificadores y aunado a ello también se puede transformar este modelo para que, con una simple escala de valores cualquier persona pueda saber si el cliente es moroso o no.

Un punto importante sobre el modelo de regresión logística es que se mencionó al inicio de la tesis por el gran uso que se tiene en las instituciones financieras y es fácil de intuir que tiene esta gran fama por obtener predicciones aceptables con un entendimiento fácil para quien lo elabora y también para el momento de presentarlo a las personas que no tienen un conocimiento solido en estadística.

Sin duda alguna los modelos de calificación crediticia son de vital importancia para las instituciones financieras, ya que gracias a estos clasificadores se puede discriminar a los malos clientes de los buenos y así obtener una tasa de impago de clientes lo más baja posible, cabe mencionar que si alguna institución no aplicase este tipo de clasificadores estaría muy expuesto al riesgo de crédito, ya que, al aceptar a un cliente seria como lanzar una moneda al aire, lo cual se reduce a un 50% de que si pague contra otro 50% de que no pague y esto ocasionaría que la institución tuviera fuertes pérdidas.

Anteriormente se ha mencionado en más de una ocasión la importancia que tiene la calificación crediticia en las instituciones financieras, pero ¿qué dificultades puede tener al implementarse este tipo de modelos? La institución se puede enfrentar a diversos riesgos, de entrada, los datos a recolectar deben de ser totalmente verídicos, en ocasiones las personas que registran al usuario pueden llegar a equivocarse o el mismo cliente puede mentir acerca de su situación actual y si no existieran las suficientes validaciones los datos introducidos al modelo no serían los correctos y ocasionaría que el clasificador en la práctica no obtenga buenas predicciones.

Otro tipo de adversidad para el desarrollo del clasificador es en el mantenimiento de este ya que en ocasiones la población puede ser afectada en un cierto ramo y provoque que las personas que antes eran buenos clientes se conviertan en morosos o viceversa, un ejemplo de lo comentado anteriormente es cuando una empresa tiene problemas económicos y ocasione que no pueda pagar la nómina completa a sus trabajadores y los afecte de tal manera que ya no puedan pagar su tarjeta de crédito.

Durante la elaboración de los clasificadores el paso más sencillo fue programarlo por el gran soporte que hoy en día otorgan los lenguajes de programación, sin embargo, el problema en el que me enfrenté fue para su comprensión ya que a pesar de tener bases sólidas en estadística

me encontré con varios métodos relativamente nuevos para mí, los cuales concluyeron en horas de estudio y dedicación.

El perfil de un actuario fue perfecto para esta tesis ya que desde el capítulo 1 se analizó el tema con una perspectiva financiera, pero a lo largo del trabajo se abordaron los diferentes temas con perspectivas estadísticas y aunado a ello, el actuario también debe de tener la capacidad de dar a entender su conocimiento a personas que no están especializadas en el área.

Anexo

Scripts de clasificadores

En esta sección se explicaron los scripts de Python (spyder) utilizados para este trabajo, tomando en cuenta que cada uno se segmenta en cinco, la primera parte hace referencia a importar las librerías y las funciones a ocupar, la segunda a la carga del archivo en donde se encuentran los datos, la tercera a la creación de las variables independientes y la dependiente y la penúltima a la segmentación (entrenamiento y de prueba) de estos y la última al código utilizado para la creación del clasificador.

```
 8  # Importamos las librerias
 9  from sklearn.model_selection import train_test_split
10  import pandas as pd
11
12  # Cargamos la base de datos y soleccionamos solo el 80% de los datos
13  bd = pd.read_csv('default of credit card clients.csv')
14
15  #Seleccionamos las variables
16  X= bd.loc[:,bd.columns != 'Y']
17  y= bd.Y
18
19  #Segmentamos a la base de datos (Entrenamiento,Prueba)
20  X_train, X_test, y_train , y_test =train_test_split(X,y,test_size=0.2,random_state=23)
```

Del código anterior tenemos que en las líneas 9 y 10 se cargaron las librerías, en la 13 a la variable “bd” se le asignó el archivo para trabajar, en las 16 y 17 se asignaron las variables independientes y la depenedientes respectivamente y en la 20 se utilizó la función train_test_split, la cual según Scikit-Learn (2021) hace referencia a “Dividir matrices o matrices en subconjuntos de prueba y tren aleatorio” , en resumen sirve para dividir a la base de datos y los argumentos que se ocuparon fueron “X,y,test_size=0.2 , random_state=23” el primer argumento hace referencia al Data Frame con variables independientes, el segundo para la variable dependiente, el tercero para el tamaño del conjunto de prueba que se obtendrá a partir de los dos primeros argumentos, el 0.2 significa que el 20% de los datos ingresados serán utilizados para probar al modelo y por último el random_state es la semilla utilizada para revolver los datos antes de dividirla, cabe recalcar que escogí el número 23 de forma arbitraria, más no tiene nada de especial.

Script regresión logística

Retomando lo comentado en el capítulo de Resultados, tenemos que en este modelo se omitió la variable **%Pago**, la cual se puede ver reflejado en la línea 18 del código presentado a continuación.

Para armar el clasificador Logit se utilizó la librería statsmodels, el cual es “statsmodels es un módulo de Python que proporciona clases y funciones para la estimación de muchos modelos estadísticos diferentes, así como para realizar pruebas estadísticas y exploración de datos estadísticos.” (Seabold, Skipper, and Josef Perktold, 2010).

Los únicos parámetros agregados a la función logit fue la variable dependiente y las independientes, ambas del conjunto de entrenamiento, después se ajusta el modelo a los datos y por último imprimí el resumen para obtener el Nova del modelo. Cabe recalcar que para usar este modelo con intercepción se agregó una columna de 1, lo cual se puede ver reflejado en la línea 25.

```
7   # Se importan las librerias que se van a ocupar
8   from sklearn.model_selection import train_test_split
9   import statsmodels.api as sm
10  import pandas as pd
11  import numpy as np
12
13  # Cargamos la base de datos y soleccionamos solo el 80% de los datos
14  bd = pd.read_csv('default of credit card clients.csv')
15
16  # Seleccionamos las variables
17  X= bd.loc[:,bd.columns != 'Y']
18  X=X.loc[:,X.columns != '%PAGO']
19  y= bd.Y
20
21  # Segmentamos a la base de datos (Entrenamiento,Prueba)
22  X_train, X_test, y_train , y_test =train_test_split(X,y,test_size=0.2,random_state=23)
23
24  # Realizamos el modelo de regresion logistica
25  X_train=sm.add_constant(X_train,prepend=True)
26  modelo=sm.Logit(y_train,X_train)
27  modelo=modelo.fit()
28  print(modelo.summary())
```

Script redes neuronales

Para este modelo se utilizó la librería keras, para importar las funciones Dense y Sequential, las cuales sirven para elaborar el clasificador de redes neuronales.

De primera instancia se tiene que en sequential es donde se van a estar agrupando las capas neuronales y se aprecia en la línea 26, después en la siguiente línea de código hace referencia a la capa oculta, la cual está conectada con las 7 variables de entrada, en esta capa ooculta existen 16 neuronas con la función de activación relu y por último se tiene la capa de salida con una neurona con la función de activación sigmoid.

```
8  # Se importan las librerias que se van a ocupar
9  from sklearn.model_selection import train_test_split
10 from keras.layers.core import Dense
11 from keras.models import Sequential
12 import pandas as pd
13
14 # Cargamos la base de datos y soleccionamos solo el 80% de los datos
15 bd = pd.read_csv('default of credit card clients.csv')
16
17 #Seleccionamos las variables
18 X= bd.loc[:,bd.columns != 'Y']
19 y= bd.Y
20
21 #Segmentamos a la base de datos (Entrenamiento,Prueba)
22 X_train, X_test, y_train , y_test =train_test_split(X,y,test_size=0.2,random_state=23)
23 X_val,X_test,y_val,y_test = train_test_split(X_test,y_test,test_size=0.5,random_state=23)
24
25
26 model = Sequential()
27 model.add(Dense(16, input_dim=7, activation='relu'))
28 model.add(Dense(1, activation='sigmoid'))
29
30 model.compile(loss='binary_crossentropy',
31               optimizer='adam',
32               metrics=['binary_accuracy'])
33
34 model.fit(X_train, y_train, epochs=30,validation_data=(X_val,y_val))
```

En las líneas de código 30 a 32 se compiló el modelo y se utilizó como función de perdida binary crossentropy, como optimizador Adam y la métrica binary_accuracy.

Por último, se ajustaron los datos al modelo utilizando los datos entrenamiento con 30 iteraciones.

Script árbol de decisión

Se ocupó la librería sklearn para importar la función tree, la cual sirve para elaborar arboles de decisión y los criterios utilizados fueron los siguientes; en criterion se optó por entropy, esta función sirve para medir la calidad de las divisiones, en min_samples_split fueron dos y es el mínimo de muestras necesarias para dividir un nodo, para min_samples_leaf son dos y es el número mínimo de muestras necesarias para estar en una hoja, max_depht es igual a tres y es la profundidad máxima que tiene el árbol y por último se tiene que splitter es igual a best y hace referencia a la estrategia utilizada para elegir la mejor división.

```
# Se importan las librerias a ocupar
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn import tree
import pandas as pd

# Cargamos la base de datos y soleccionamos solo el 80% de los datos
bd = pd.read_csv('default of credit card clients.csv')

# Seleccionamos las variables
X= bd.loc[:,bd.columns != 'Y']
y= bd.Y

# Segmentamos a la base de datos (Entrenamiento,Prueba)
X_train, X_test, y_train , y_test =train_test_split(X,y,test_size=0.2,random_state=23)

# Crear Arbol de decision con profundidad = 3
algoritmo = tree.DecisionTreeClassifier(criterion='entropy',
                                        min_samples_split=2,
                                        min_samples_leaf=2,
                                        max_depth = 3, splitter='best')
algoritmo.fit(X_train, y_train)

```

Script análisis discriminante lineal

De la librería sklearn se importó la función LinearDiscriminantAnalysis y se le abrevió como LDA, esta función nos permitió elaborar con facilidad el clasificador de discriminante lineal, como se aprecia en la línea de código 26 el único atributo que tiene el modelo es el de solver, del español solucionador y hace referencia a la función que se ocupa para estimar a las betas, dicho parámetro se optó por usar lsqr que significa mínimos cuadrados.

Por último, en la línea 27 se le ajustó al modelo los datos de entrenamiento.

```
8   #                   Importamos las librerias
9   from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
10  from sklearn.model_selection import train_test_split
11  from sklearn.metrics import confusion_matrix
12  from sklearn.metrics import accuracy_score
13  import pandas as pd
14
15  #                Cargamos la base de datos y soleccionamos solo el 80% de los datos
16  bd = pd.read_csv('default of credit card clients.csv')
17
18  #Seleccionamos las variables
19  X= bd.loc[:,bd.columns != 'Y']
20  y= bd.Y
21
22  #Segmentamos a la base de datos (Entrenamiento,Prueba)
23  X_train, X_test, y_train , y_test =train_test_split(X,y,test_size=0.2,random_state=23)
24
25  #Metodo discriminante
26  lda = LDA(solver='lsqr')
27  lda.fit(X_train,y_train)
```

Referencias

Andreola, R. y Haertel V. (2010). *Classificação de imagens hiperespectrais empregando support vector machines* [Clasificación de imágenes hiperespectrales utilizando máquinas de vectores de soporte].

Banxico. (2005). *Definiciones básicas de riesgo*.

Basilea. (1999). *Principios para la administración del riesgo de crédito*. Comisión de Basilea de supervisión de bancos.

Basilea. (s.f.). *Principios para la administración de riesgos de crédito*. Obtenido de http://www.asbasupervision.com/es/todos/biblioteca-virtual-asba/gestion-de-riesgos/riesgo-de-credito/144-gr-rc03/file

Cuartas Aguirre, F. (2013). *Banca comercial y de inversión*. Bogotá.

de Lara Haro, A. (2008). *Medición y control de riesgos financieros*. México D.F.: Limusa S.A DE C.V.

del Valle Benavides, A. R. (s.f.). *Curvas ROC (Receiver-Operating-Characteristic) y sus aplicaciones* [Trabajo final para el grado de matemáticas, Universidad de Sevilla].

Española, R. A. (2020). *Diccionario de la Lengua Española*. Obtenido de https://dle.rae.es/cr%C3%A9dito

Gómez, F. d. (24 de marzo de 2014). *Las 5 c's del crédito: blog UDLAP*. Obtenido de http://blog.udlap.mx/blog/2014/03/las5cdelcredito/

Gonzalo, B. H. (1996). *El sistema financiero en México*.

Henríquez Muñoz, C. N. (2014). *Estudio de Técnicas de análisis y clasificación de señales EEG en el contexto de Sistemas BCI (Brain Computer Interface)* [Tesis de Maestría, Universidad Autónoma de Madrid].

Jacques Marcuse, R. (2009). *Diccionario de términos financieros y bancarios*. Ecoe Ediciones.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R* [Introducción al aprendizaje estadístico con aplicaciones en R] Springer.

Lara Albín, J. D. (2014). *Técnicas de clusterización*. Obtenido de http://bibing.us.es/proyectos/abreproy/5453/fichero/PFC+tecnicas+clusterizacion.pdf

Larranaga, P., Inza, I. & Moujahid, A. (s.f.). *Redes Neuronales*.

Leroy Miller, R. (1992). *Moneda y banca*. Bogotá: Mc Graw Hill.

Lizárraga Mollinedo, C. (s.f.). *El Índice de Gini: la desigualdad a la palestra*.

Morales Castro, J. A. & Morales Castro, A. (2015). *Crédito y cobranza*. México D.F.: Grupo Editorial Patria.

Páez juka, s. d. (2019). *Análisis comparativo de herramientas open source para data mining sobre datos públicos del ministerio de educación de la república del ecuador.*

Saavedra García, M. L. & Saavedra García, M. J. (2010). *Modelos para medir el riesgo de crédito de la banca*. Obtenido de Scielo: http://www.scielo.org.co/pdf/cadm/v23n40/v23n40a13.pdf

Scikit-Learn. (2020). *Neural network models (supervised)* [Modelo de redes neuronales (supervisado)] Obtenido de https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Valle Carrascal, J. M. (2015). *Modelos de medición de riesgo de crédito* [Tesis doctoral, Universidad Complutense de Madrid].

Valores, C. N. (2005). *Disposiciones de carácter general aplicables a las instituciones de crédito*. Obtenido de https://www.cnbv.gob.mx/Prensa/Presentaciones%20Seminario%20Corresponsales/i.%20Circular%20%C3%9Anica%20de%20Bancos.pdf

Valores, C. N. (2013). *Preguntas frecuentes*. Obtenido de https://www.cnbv.gob.mx/SECTORES-SUPERVISADOS/BANCA-MULTIPLE/Paginas/Preguntas-Frecuentes.aspx

Vandemaele, S., Vergauwen, P. & Michiels, A. (2009). *Management Risk Reporting Practices and their determinants* [Prácticas de gestión de informes de riesgos y sus determinantes].

Zorrilla Arena, S. (1994). *Diccionario de Economía*. México: Limusa