ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE SISTEMAS

UNIDAD DE TITULACIÓN

DESARROLLO DE UN MODELO DE PREDICCIÓN BASADO EN ALGORITMOS DE MACHINE LEARNING PARA MEDIR EL RIESGO CREDITICIO.

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MAGISTER EN SISTEMAS DE INFORMACIÓN

OSCAR IVAN PUCHA GUALOTO

oscar.pucha01@epn.edu.ec

Director: Sang Guun Yoo, Ph.D

sang.yoo@epn.edu.ec

2022

APROBACIÓN DEL DIRECTOR

Como director del trabajo de titulación “Desarrollo de un modelo de predicción basado en algoritmos de machine learning para medir el riesgo crediticio”, elaborado por Oscar Iván Pucha Gualoto, estudiante de la Maestría en Sistemas de Información habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la Defensa oral.

Firmado electrónicamente por:
SANG GUUN YOO .

Sang Guun Yoo, Ph.D
DIRECTOR

DECLARACIÓN DE AUTORÍA

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

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

Oscar Iván Pucha Gualoto

DEDICATORIA

Al ser que cada día llena mi vida de amor, fortaleza y alegría motivándome en los momentos difíciles y levantando mi espíritu con sus palabras y acciones.

Tania mi amada esposa.

A mis queridos hijos Zack y Abner, razón de mi esfuerzo, pasión y felicidad; a mis padres ejemplo de dedicación y a mis hermanos por su compañía y aprecio.

AGRADECIMIENTO

A Dios el dador de la vida, a mi familia por su comprensión y paciencia durante estos años dedicados a la instrucción universitaria.

A la Escuela Politécnica Nacional y a su planta docente por brindarme nuevos conocimientos y capacidades que benefician mi calidad profesional.

Un agradecimiento especial a mi tutor Sang Guun Yoo, Ph.D, por su indispensable e incondicional apoyo en la realización del presente trabajo.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

INDICE DE ANEXOS

RESUMEN

El presente documento tiene como objetivo central, plantear un modelo de predicción de riesgo de crédito basado en el uso de machine learning, con el fin de ser empleado en la evaluación crediticia en instituciones comerciales y financieras. De esta manera, se pretende minimizar la probabilidad de entregar un crédito a una persona que no tenga la capacidad suficiente para resarcir el dinero entregado por el banco, cooperativa o cualquier otro tipo de institución del sistema financiero o comercial. La metodología empleada tiene un carácter mixto, porque se emplearon herramientas de análisis cualitativo y cuantitativo. Para sistematizar las actividades para el desarrollo del modelo, se emplearon metodologías agiles. Los modelos de machine learning escogidos fueron Regresión Logística, Árboles de Decisión, Vecino más próximo y Support Vector Machine. La conclusión general del documento se asocia a que todos los modelos empleados arrojan resultados satisfactorios y pueden ser usados para la evaluación del riesgo crediticio. Los modelos propuestos para cumplir los objetivos planteados que han mostrado el mejor ajuste en sus resultados fueron la Regresión logística y el Support Vector Machine. Es tarea de las instituciones financieras, evaluar el modelo que mejore se adapte a sus procesos crediticios y tenga mejor relación con sus objetivos de negocio; ya que el modelo de predicción empleado es una herramienta de apoyo a la gestión comercial con el fin de precautelar el patrimonio y rentabilidad de la institución financiera y no se constituyen en un mecanismo de presión al cliente para mejorar la cartera de crédito.

ABSTRACT

The main objective of this document is to propose a credit risk prediction model based on the use of machine learning, to be used in credit evaluation in financial or commercial institutions. In this way, it is intended to minimize the probability of giving a loan to a person who does not have sufficient capacity to repay the money given by the bank, cooperative or any other type of institution in the financial and commercial system. The methodology used has a mixed character, because qualitative and quantitative analysis tools were used. Agile methodologies were used to systematize the activities for the development of the model. The chosen machine learning models were Logistic Regression, Support Vector Machine, Decision Tree and Nearest Neighbor. The general conclusion of the document is associated with the fact that all the models used give satisfactory results and can be used for risk prediction. The models proposed to meet the stated objectives that have shown the best fit in their results were the Logistic Regression and the Support Vector Machine. However, it is the task of financial institutions to evaluate the model that best suits their credit processes and has a better relationship with their business objectives; since the prediction model used is a tool to support commercial management to safeguard the equity and profitability of the financial institution and does not constitute a mechanism to pressure the client to improve the loan portfolio.

1. INTRODUCCIÓN

1.1 Planteamiento del problema

El problema de investigación del presente documento se concentra en las organizaciones que tienen como fin alternativo de otorgar créditos. Estas organizaciones suelen presentar deficiencias en el análisis de riesgo de sus clientes; este es un caso que sucede con relativa frecuencia en empresas comerciales que se enfocan en negocios B2C (business to customer) o empresas industriales y comerciales que se enfoca en B2B (business to business).

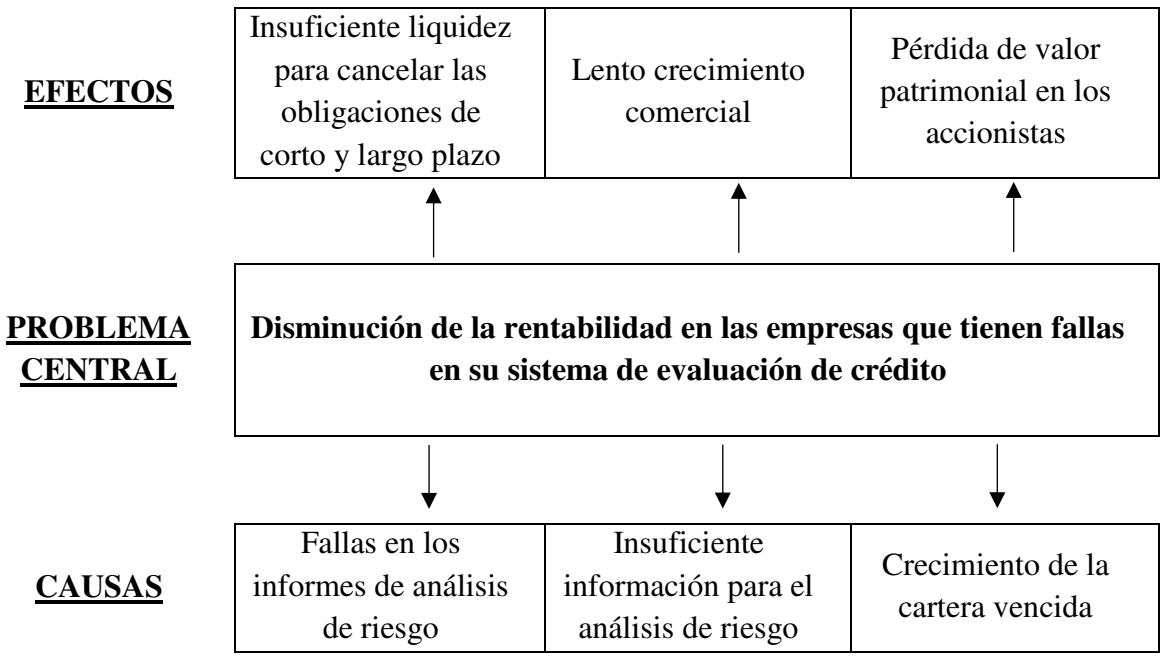
Esta deficiencia en las organizaciones conlleva a errores en la toma de decisiones, debido a fallas en la comunicación entre los participantes en el análisis, insuficiente documentación o errores en los informes crediticios (Borrero & Bedoya, 2020); estas fallas incrementan el riesgo crediticio en las empresas y por ende ponen en riesgo los indicadores de liquidez, que en el mediano y largo plazo se transforma en una situación de baja rentabilidad; es por ello, que este trabajo considera profundizar el aspecto relacionado con el análisis de crédito en las organizaciones.

Esta investigación se encamina en estudiar el campo especifico del riesgo de crédito, el cual se refiere a la posibilidad de que un cliente no cumpla con sus obligaciones adquiridas (Cela & Cuenca, 2018), es por ello, que el estudio del riesgo de crédito constituye en una fuerte herramienta de análisis para las organizaciones que tienen como fin en su modelo de negocio otorgar financiamiento a sus clientes, ya que permite anticipar posibles pérdidas por no pago y la posterior generación de una cartera vencida que afecte la liquidez de la organización.

El proceso de evaluación de crédito considera criterios que se asocian a la capacidad de pago, tomando en cuenta el comportamiento del solicitante de crédito, para lo cual captura información proveniente de datos cuantitativos como la información de la central de riesgos, estados contables o declaraciones de impuestos; adicionalmente, se utiliza datos cualitativos como entrevistas o informes de las personas que mantienen un contacto directo con el solicitante del crédito (Romillo, 2019).

Tomando en cuenta los aspectos mencionados en párrafos anteriores, se plantea el siguiente árbol del problema que contiene las causas y efectos que orientan el presente documento:

Figura 1. Árbol del problema



La figura anterior muestra la relación que existe entre las fallas en el análisis de riesgo crediticio, el cual desencadena en una falla estructural en la organización que presente estos problemas, la cual puede desembocar en una disminución en la rentabilidad, debido a que existen falencias en la liquidez, bajo crecimiento comercial asociado a la poca disponibilidad de recursos y pérdida de valor patrimonial para los accionistas (Ross, Westerfield, & Jaffe, 2014).

En vista de aquello, es necesario que las organizaciones asuman que la calificación del riesgo crediticio es un proceso clave para lograr un posicionamiento competitivo en la industria donde desarrollen sus actividades comerciales, caso contrario se enfrentan a fuertes incertidumbres que se asocian con pérdidas financieras y la asunción de procesos legales asociados a la recuperación de cartera vencida.

En el caso de continuar con este escenario de fallas en el análisis de riesgo y posterior entrega de crédito, las empresas se encontrarían ante un panorama de incertidumbre en el aspecto comercial, operativo y financiero, que pondría en peligro la sustentabilidad de la empresa en el mediano y largo plazo.

1.2 Objetivo general

Desarrollar un modelo de predicción basado en algoritmos usado de machine learning para medir el riesgo de crédito y la entrega de monto crediticio.

1.3 Objetivos específicos

- Cimentar los aspectos teóricos que respaldan el desarrollo del modelo machine learning para predecir el riesgo de crédito en una operación comercial.

- Analizar la información que forma parte del modelo machine learning para predecir el riesgo de crédito.

- Evaluar los componentes del modelo de predicción de riesgo de crédito en base al uso de los algoritmos de machine learning.

1.4 Justificación

Ante la demanda creciente por financiamiento, las empresas tienen la necesidad de demandar procesos asociados a la calificación de crédito, mediante modelos de riesgo crediticio; estos modelos se enfocan en la predicción de resultados en el incumplimiento de pago.

El tema expuesto en el presente documento pretende investigar las opciones que tienen las empresas para asumir mejores procesos de evaluación de riesgo de crédito, una de las alternativas que tiene mejor exposición, es la asunción de la tecnología como mecanismo para minimizar el impacto del no pago en una operación comercial.

En lo que corresponde a la justificación teórica, este proyecto se ampara en los mecanismos tecnológicos, específicamente en las bondades de la inteligencia artificial, como herramienta de predicción, con el uso de algoritmos de machine learning. Byanjankar (2017) menciona que el uso de algoritmos de aprendizaje automático son modelos basados en datos de muestra, con el fin de programar tareas automatizadas.

Los algoritmos usados en la investigación deben enfocarse en el riesgo de no pago, para determinar que cliente estaría asociado a una situación de no cancelar el valor asignado en

su crédito. El segundo enfoque, se asocia al riesgo de crédito, donde el modelo debe decidir si se entrega el crédito solicitado (Borrero & Bedoya, 2020).

De esta manera, el modelo tiene un enfoque global en el proceso de calificación y análisis de riesgo de crédito; para construir estas aristas del modelo, existe la suficiente evidencia empírica, que permiten fundamentar la realización del presente documento.

Se puede decir, que existe un sin número de modelos que se han utilizado en la evaluación del riesgo crediticio, entre los que más destacan están los modelos de *credit scoring* los mismos que plantean la automatización del proceso de gestión de créditos, en cuanto a asignar o no un crédito personal, el mismo que está basado a un conglomerado de características notables que permitan la toma de la decisión. Lo interesante de la aplicación de este modelo se basa en efectividad del algoritmo utilizado y del apoyo de un sistema eficaz de análisis de datos. Además, existen los modelos relacionales, los mismos que permiten el análisis de la información histórica del solicitante, valorando el riesgo crediticio que puede tener el cliente al efectuar futuras operaciones comerciales. No obstante, este proceso excluye a los clientes nuevos, debido a la falta de información histórica. También existen los modelos de evaluación económico - financiero, los mismos se enfocan en la evaluación de los estados financieros y de la industria. Debido a esto, se plantea la construcción de un modelo que haga uso de herramientas de vanguardia y actualidad (i.e. inteligencia artificial, aprendizaje automático) que permitan ofrecer a este sector empresarial, un mecanismo de selección de clientes potenciales minimizando el riesgo crediticio y por ende la posibilidad de pérdidas o de costosas querellas.

La justificación práctica se lleva a cabo con la puesta en marcha del modelo de riesgo basado en algoritmos de inteligencia artificial, que obtiene una mejor toma de decisiones en la entrega de crédito en las organizaciones, minimizando el impacto de la presencia de cartera vencida, la cual puede afectar directamente en la rentabilidad de la empresa; de acuerdo a lo mencionado por Romillo (2019), este proceso asertivo y predictivo permite minimizar las perdidas asociadas a cartera no cobrada o en procesos judiciales.

Esta investigación tiene como beneficiarios directos a las organizaciones y sirve para mejorar los procesos de concesión de crédito, ya que con la puesta en marcha de este modelo en su proceso de concesión minimizan el impacto asociado a riesgo de no pago. Es relevante la investigación por cuanto se enfoca en procesos estratégicos y de apoyo en la organización, proveyendo herramientas basadas en datos de muestra que sirven para la correcta toma de decisiones.

Es factible la aplicación del presente trabajo, puesto que la información es accesible y apta para aplicarla con la metodología de machine learning, ya que toma datos que reportan los clientes para acceder al crédito o están disponibles en el buró de crédito.

1.5 Metodología

El presente trabajo considera un enfoque mixto, que consiste en el uso de técnicas cualitativa con la formulación de entrevistas a expertos y cuantitativa con la recolección de datos muestrales; el tipo de investigación es descriptiva porque busca narrar hechos y circunstancias en el desarrollo del modelo de machine learning. El método que se utiliza es deductivo y se utiliza fuentes de datos primarios con la entrevista y datos muestrales y fuentes

secundarias con el uso de otras investigaciones especializadas, revistas y textos enfocados en machine learning.

El desarrollo de un modelo de machine learning es una tarea que debe asociarse a una metodología de gestión y desarrollo de proyectos de software, para lo cual se ha escogido la metodología ágil, que se enfoca en maximizar los equipos de trabajo mediante el uso de recursos como el tiempo y el presupuesto como aspectos claves para el diseño del proyecto. La metodología ágil, define al proyecto como una misión temporal que se aplica para lograr un grupo de objetivos en un período especifico de tiempo, por tanto, un proyecto debe tener objetivos, tiempos y presupuesto de ejecución (McCarthy, 2020). Los cuatro pilares del desarrollo de metodologías ágiles son los siguientes (véase la **Figura *2*Error! Reference source not found.Error! Reference source not found.**):

a. Personas e interacciones: la metodología otorga mayor valor a las personas e interacciones sobre las herramientas y los procesos, ya que las personas ofrecen soluciones a problemas que pueden llegar a suceder en el desarrollo de software, esto permite una mejor adaptación al cambio.
b. Software funcional: este pilar busca simplificar el uso de documentación y enfatiza en la creación de las historias del usuario, donde se transmite las necesidades que requiere el diseño del proyecto.
c. Colaboración con el cliente: en el desarrollo del proyecto es necesario escuchar al cliente para conocer sus requerimientos reemplazando la colaboración en lugar de la negociación.
d. Respuesta ante el cambio: la metodología exige responder al cambio en lugar de evitarlo, el proyecto requiere planes elaborados que en ocasiones se desvían de su

camino inicial, lo cual implica un mayor gasto; por ello la metodología ágil propone adaptarse al cambio.

Figura 2. Valores Metodología Ágil

Fuente: (McCarthy, 2020)

A continuación, se presenta una descripción detalla de la metodología para cada objetivo:

- **Objetivo 1:** Fundamentar los aspectos teóricos que respaldan el desarrollo del modelo machine learning para predecir el riesgo de crédito en una operación comercial.

El cumplimiento de este objetivo tiene como base a la investigación teórica, debido a permite la revisión e indagación del desarrollo, la conceptualización, características, objetivos, ventajas, limitaciones y elementos del modelo machine learning para predecir el riesgo de crédito.

Según Bernal (2018) la investigación teórica es la revisión conceptual de una teoría o de alguna de sus partes o aspectos, el contratarla, comprobarla, validarla o verificarla, debatirla y contradecirla.

- **Objetivo 2:** Analizar la información que forma parte del modelo machine learning para predecir el riesgo de crédito.

Con el fin de cumplir con este objetivo, se aplicará la investigación de enfoque mixto, con la recepción de información a través de entrevistas con expertos y la recolección de datos muestrales, que sirven de fundamento para el planteamiento de los algoritmos y la posterior construcción del modelo de machine learning (Rodriguez & Miñano, 2017).

Las fuentes primarias permitirán la recopilación de información relevante para el alcance del segundo objetivo, en este caso la entrevista dirigida a expertos, que busca conocer las necesidades de las empresas en cuanto a la calificación de riesgo crediticio y las expectativas que buscan para asumir un modelo de estas características en su empresa.

De acuerdo con Gómez (2014), la entrevista es una forma de recopilar información con la participación directa de los involucrados, su objetivo es brindar un enfoque amplio de los factores y características que inciden en la investigación.

La entrevista dirigida será el medio a través del cual se obtenga la información requerida para el desarrollo del presente trabajo, para lo cual se aplicará un cuestionario que tendrá una duración de máximo 10 minutos logrando con ello disminuir la cantidad de cuestionamientos sin respuesta, además de que permite la adquisición de información relevante y concreta.

Objetivo 3: Evaluar los componentes del modelo de predicción de riesgo de crédito en base al uso de algoritmos de machine learning.

Para alcanzar este objetivo, se toma en cuenta el tipo de investigación descriptiva, que tiene el propósito de especificar los componentes del modelo, tomando como referencia la

recolección de información de parte del segundo objetivo y la fundamentación teórica del primer objetivo. El método deductivo se utilizará para el alcance del tercer objetivo, mismo que va de lo general (que es la información captada en el primer objetivo) a lo particular (información captada en el segundo objetivo), con estos pilares se construye el modelo que constituye la solución al problema planteado.

2. MARCO TEÓRICO

2.1 Antecedentes de la inteligencia artificial

El inicio formal de la inteligencia artificial tuvo su asidero en los trabajos del científico Alan Turing en la década de 1950, su investigación estuvo centrada en los preceptos de la lógica, neurociencia y las matemáticas, buscando la respuesta a una duda razonable, respecto a que si las máquinas podían pensar por sí mismas. Los antecedentes de la inteligencia artificial inician antes de la aparición de los computadores o la electrónica, el ser humano siempre sintió la necesidad de crear inteligencia fuera del cerebro. Desde la escritura de la sociedad griega antigua ya se conocían claros deseos de los filósofos de organizar pensamientos e ideas alrededor de máquinas que alcancen el razonamiento humano. En el Medievo, los árabes tuvieron esfuerzos notables por crear maquinas pensantes.

Entre el siglo XVI y siglo XIX existieron varios intentos de crear máquinas para actividades diversas como el ajedrez, ayuda en la medicina y para crear música. Fue hasta el siglo XX, en la vigencia de la Segunda Guerra Mundial, cuando se crearon las primeras máquinas de inteligencia artificial para descifrar mensajes encriptados, fueron los británicos y estadounidenses al mando de Turing los primeros en construir un ordenador mecánico. En 1950, este mismo científico escribe su tratado sobre computadores e inteligencia que es considerado un auténtico tratado del tema, por lo que, es considerado el padre de la inteligencia artificial (Delgado, 2016).

La información expuesta permite identificar dos aspectos básicos, la inteligencia artificial siempre estuvo rondando la mente del ser humano, mucho antes de la aparición de los computadores y el hecho de que la inteligencia artificial no tuvo un único punto de partida,

sino que fue una suma de esfuerzos de científicos y mentes que se atrevieron a pensar más allá de la realidad.

El punto de partida formal de la inteligencia artificial sucede en la Conferencia de Dartmouth, en el verano de 1956 se reunieron científicos de las áreas de matemáticas, neurología, psicología e ingeniería eléctrica, el punto en común de los participantes era el uso de computadores para diferentes actividades, el objetivo era normalizar y coordinar esfuerzos para crear actividades que regulen sus propuestas; de esta manera nace la informática con el aporte de diferentes ramas de la ciencia.

Esta conferencia tiene alta importancia para las áreas del conocimiento detrás de la inteligencia artificial, porque estableció los pilares para la promover la investigación y orientar los esfuerzos de los académicos por desarrollar una hoja de ruta para impulsarla, se puede mencionar los siguientes pilares:

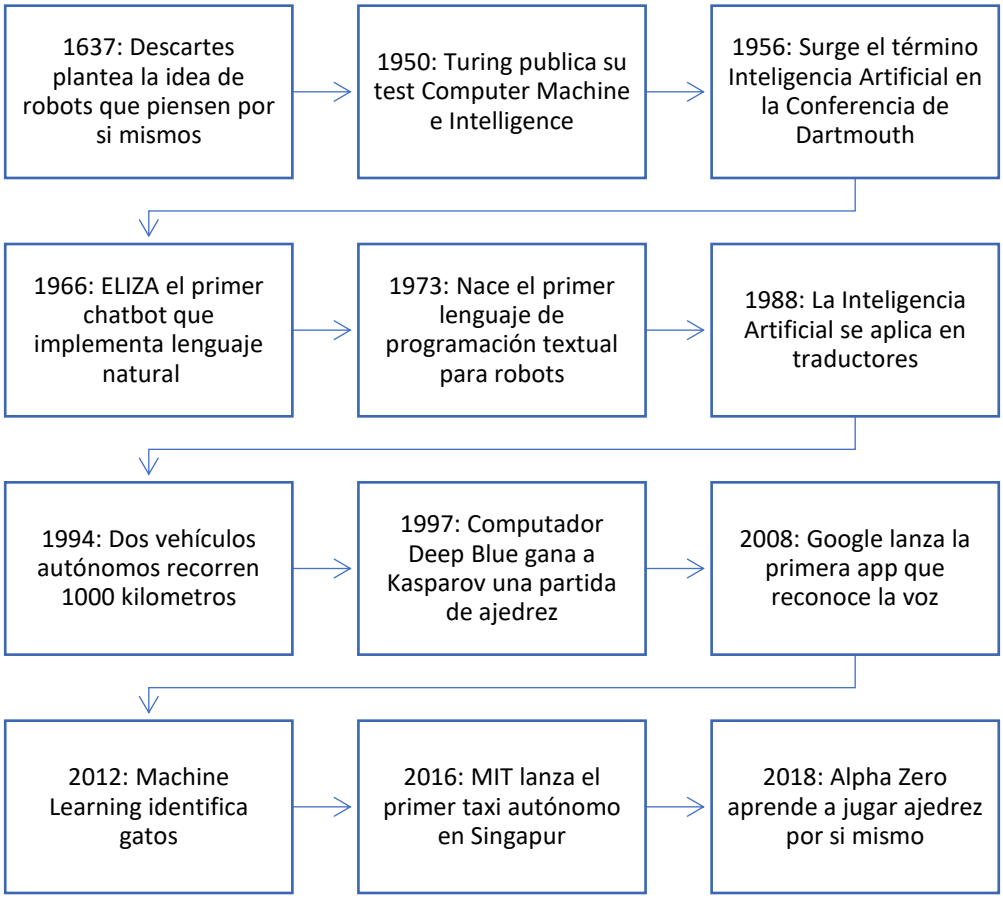
1. Computadores automáticos: la capacidad del computador está directamente relacionada con su habilidad para el comportamiento inteligente, mientras mayor capacidad tenga un computador mayor capacidad para implementar inteligencia artificial.
2. Programación usando lenguaje natural: permite trasladar la premisa que el lenguaje es el punto de partida del pensamiento humano, por lo que, se debe trasladar este aspecto a los computadores.
3. Redes neuronales: organizar las ideas como un cerebro humano mediante el uso de redes comunicacionales.

4. Teoría del tamaño del cálculo: es necesario incrementar la capacidad de cálculo de los computadores.
5. Auto mejora: los computadores deben tener la capacidad para mejorar sus propios procesos.
6. Abstracción: investigar, clasificar y describir los métodos para que una máquina pueda procesar datos sensoriales.
7. Aleatoriedad y creatividad: guiar los pensamientos autónomos de un computador para mejorar su aprendizaje.

La puesta en marcha de estos pilares permitió el desarrollo de la inteligencia artificial hasta la capacidad de aprendizaje que tiene en la actualidad, en este aspecto radica la importancia de la Conferencia de Dartmouth como el puntapié inicial para una ciencia capaz de revolucionar la interacción entre máquinas y humanos.

Como ordenamiento de las ideas sobre la inteligencia artificial y su evolución se expone los siguientes hitos que influyeron en su desarrollo y aspectos importantes que se debe tomar en cuenta para comprender hacia donde se dirige la ciencia en el futuro:

Figura 3. Hitos de la inteligencia artificial



Fuente: (Bidaurrazaga, 2019)

La figura anterior muestra la evolución de la inteligencia artificial, que ha tenido un crecimiento constante desde la década de 1980, con un sinnúmero de aplicaciones en actividades donde antes solo podía intervenir el ser humano, en algunos casos ha mejorado su participación, tornando como territorio exclusivo de los computadores, como es el caso del análisis para la toma de decisiones en campos desde la medicina hasta las finanzas.

2.2 Inteligencia artificial

Turing definía a la inteligencia artificial como una de las ramas de las ciencias computacionales capaz de desarrollar herramientas racionales para la resolución optima de

problemas (Molina, 2014). En 1968 Minsky define a la inteligencia artificial como la ciencia que hace que las máquinas hagan cosas que requieren inteligencia; en 1981, Barr y Feigenbamm definieron a la inteligencia artificial como la ciencia capaz de diseñar sistemas informáticos inteligentes, que asocien aprendizaje y razonamiento para la resolución de problemas. En 1985, H.A. Simon establece que el objetivo de la inteligencia artificial es duplicar y mejorar el pensamiento humano en una máquina capaz de asimilar una conducta inteligente (Delgado, 2016).

Rouhiainen (2018) define a la inteligencia artificial como la facultad que tienen las máquinas para utilizar los algoritmos, estudiar los datos y aplicarlos al momento de tomar una decisión, como lo hace un ser humano. La diferencia radica en que las máquinas no necesitan descanso y pueden procesar un alto volumen de información; otro punto a destacar consiste en la menor posibilidad de cometer errores en la toma de decisiones por parte de las máquinas.

En la actualidad existen muchas aplicaciones asociadas a la inteligencia artificial, algunas de ellas, forman parte de nuestra vida sin ser conscientes de su uso, en este aspecto radica precisamente su capacidad para adaptarse al comportamiento del ser humano, sin ser detectado y trabajando sigilosamente. Una exposición de las aplicaciones de la inteligencia artificial son las siguientes:

- Reconocimiento de imágenes
- Procesamiento de datos de pacientes
- Mantenimiento predictivo
- Clasificación de objetos
- Distribución de contenidos en redes sociales

- Protección contra amenazas cibernéticas

En este contexto de la inteligencia artificial, el aprendizaje automático es su principal perspectiva. El aprendizaje automático es la capacidad de los ordenadores, máquinas o programas en aprender por sí mismos, sin necesidad de estar programados para ello, un ejemplo consiste en las sugerencias o predicciones que ofrece una red social como Facebook o YouTube (Rouhiainen, 2018).

Varios autores coinciden en expresar que la inteligencia artificial es un nuevo factor de la producción, que permitirá a las empresas impulsar su crecimiento a través de la automatización inteligente y la capacidad que tengan las organizaciones de adaptar su realidad a los retos que presenta la innovación en el campo de la inteligencia artificial.

2.3 Machine Learning

El término machine learning o aprendizaje automático, se refiere a la capacidad que tiene una máquina, dispositivo o software de generar un aprendizaje propio, basado en algoritmos de programación y en la entrada de datos que recibe de su entorno. El machine learning es un avance tecnológico que permite la automatización de tareas, minimizando la intervención del ser humano.

Jones (2019) el aprendizaje automático es ensañarle a una máquina o software a realizar una tarea y que cada vez pueda realizar esta tarea de mejor manera y en menor tiempo posible, el análisis de correos electrónico como spam, es una tarea de aprendizaje automático en base de la experiencia e interacción con el usuario del correo electrónico, al realizar una acción de clasificar el correo como no deseado o spam.

Según Aceituno (2019) la conceptualización de machine learning se refiere a la habilidad de un computador que aprenda en base de la experiencia, que es generada a través de la asimilación de datos y actividades. La máquina o el software está en capacidad de predecir escenarios futuros emprender acciones de forma automática, que desencadenan en otras acciones asociadas al comportamiento del usuario o el modelo donde se aplica el aprendizaje automático.

2.3.1 Algoritmos en Machine Learning

El aspecto central del aprendizaje automático es la creación y aplicación de un algoritmo de programación, que es una secuencia de pasos y/o actividades que permiten dar solución a un problema determinado. Monasterio Astobiza (2017) define al algoritmo como una lista de instrucciones que llevan directamente a una respuesta o resultado particular, basado en la información disponible. Hill (2016) define al algoritmo como un constructo matemático con una estructura de control finita, abstracta y efectiva de acción imperativa para cumplir con un propósito definido.

El uso de los algoritmos se conoce desde la creación de los primeros tratados de algebra y cálculo en el siglo IX hasta forma en la actualidad una parte central de las ciencias computacionales, informática y diversas ramas de la ingeniería. Diversos autores han denominado a la revolución algorítmica, el uso intensivo de estas herramientas de cálculo en las ciencias computacionales, que ha tenido una fuerte incidencia en el diseño de software a partir de la primera década del siglo XXI.

De acuerdo con Tejero (2020) los algoritmos se clasifican según su sistema de signos, función y estrategia, como se muestra en la siguiente tabla:

Tabla 1. Clasificación de algoritmos

Tipo de algoritmos	Clasificación
Sistema de signos	Algoritmos cualitativos
	Algoritmos cuantitativos
	Algoritmos computacionales
	Algoritmos no computacionales
Función	Programación dinámica
	Algoritmos de ordenamiento
	Algoritmos de búsqueda
	Algoritmos de backtracking
	Branch and bound
	Algoritmo de marcaje
	Algoritmo de encantamiento
Estrategia	Algoritmos probabilísticos
	Algoritmos paralelos
	Divide & Conquer
	Algoritmos determinísticos
	Algoritmos no determinísticos
	Algoritmo greedy
	Algoritmos heurísticos
	Algoritmos cotidianos
	Algoritmos de escalada

Fuente: (Tejero, 2020)

Los algoritmos son el aspecto central del machine learning, ya que son automatizaciones que se usan para predecir resultados, la tarea de los programadores consiste en traducir los problemas cotidianos a un lenguaje que entienda una máquina; con la aparición del internet en la transmisión de información y el uso de enormes cantidades de datos (Duarte, 2018), lleva al machine learning a un espacio ideal para ser aplicado en innumerables tareas, que

dependen de la imaginación del programador y su capacidad de entender el problema que busca resolver.

Kotsiantis (2017) establece que el machine learning es el aprendizaje constante en la búsqueda de algoritmos, a partir de datos externos para producir hipótesis que posteriormente se traducen en soluciones futuras. Zhou (2017) menciona que el éxito de la aplicación de los algoritmos en el machine learning es asociado a las altas fuentes de información y datos que existen en la actualidad.

2.3.2 Tipo de Aprendizaje Automático

El campo del aprendizaje automático divide en dos tipos, según exista o no retroalimentación en el proceso, estos son:

- Aprendizaje automático no supervisado: realizan el proceso basado únicamente en las entradas, se van procesando las observaciones según la información que va recibiendo la máquina o el software (Baviera, 2016). En este tipo de aprendizaje no existe conocimiento previo de los datos del modelo, ejemplos de este tipo de aprendizaje son K-means, clustering, componentes principales e independientes.
- Aprendizaje automático supervisado: se realizan dos procesos que permiten evaluar los mejores parámetros para el aprendizaje y evalúa el nivel de fiabilidad estos parámetros (Baviera, 2016); el proceso es una especia de entrenamiento donde existe una mejora continua para adaptarse al mejor resultado posible según lo dispuesto en la parametrización del algoritmo. Ejemplos de aprendizaje automático supervisado son regresiones lineales y logísticas, redes neuronales y máquinas de vectores.

El campo de aplicación del aprendizaje automático es variado y no se concentran en una actividad específica del conocimiento humano, una aplicación diaria se encuentra en el sistema de detección de búsqueda de Google o en el sistema de reconocimiento facial del teléfono móvil. Otro campo de aplicación es la medicina con la secuencia del ADN, la creación de imágenes de alta calidad y en el análisis del comportamiento del consumidor. Esos son algunos de los ejemplos de aplicación del aprendizaje automático, no son los únicos, en el campo financiero y bancario existen modelos de predicción que ayudan a la decisión en la entrega de créditos en instituciones financiera. La capacidad de aplicación depende del programador y su formulación del problema que se busca solucionar.

2.4 Modelos de aprendizaje automático

Los modelos de aprendizaje automático pueden agruparse en tres tipos, que se detallan a continuación:

2.4.1 Modelos lineales

Este tipo de modelo buscan una línea de ajuste en los datos que dispone, el ejemplo más conocido son la regresión lineal, conocida como la regresión de mínimos cuadrados y la regresión logística que utiliza la adaptación a variables discretas. De acuerdo a Aceituno (2019) la regresión logística es uno de los modelos más utilizados en el análisis de riesgo crediticio. La capacidad de análisis de estos modelos es limitada porque existe un ajuste muy cercano a los datos, que puede representar fallas en comportamientos más complicados (Sandoval, 2018).

El modelo de regresión lineal utiliza el método de los mínimos cuadrados, que consiste en calcular la suma de las distancias al cuadrado entre los puntos reales y los puntos definidos por la recta de regresión, de modo que la mejor estimación es la que minimice las distancias entre los puntos (Pelaez, 2016).

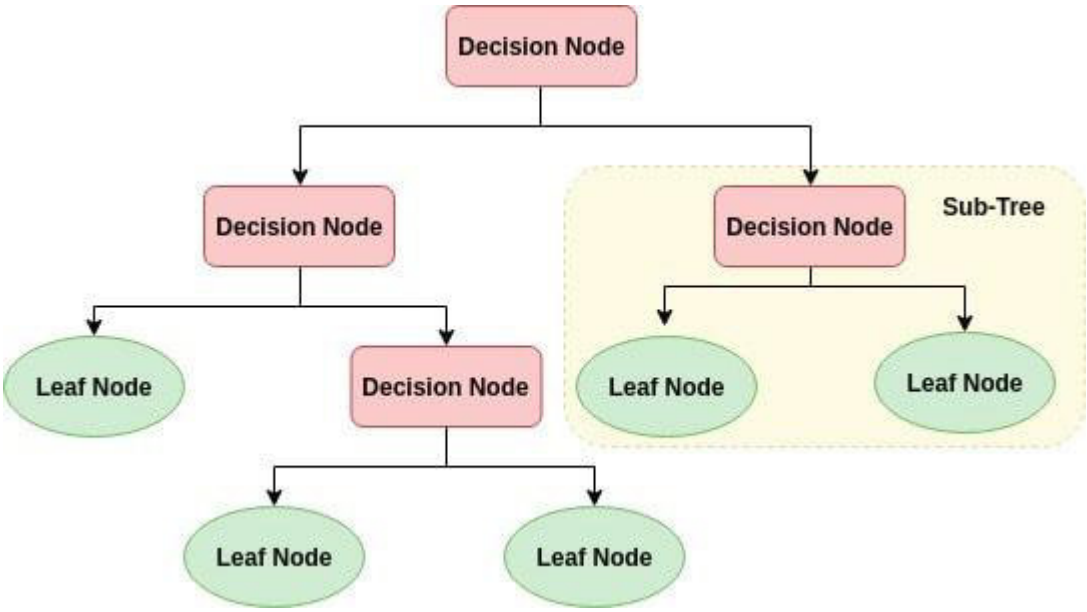
El modelo de regresión logístico es conocido como el modelo Logit, que es utilizado en diversos modelos de predicción de pago (Grau, 2020). El primer paso en la construcción del modelo es diferenciar el tipo de variable sea endógena o dependiente, así como determinar si el tipo de variable es cuantitativa o cualitativa.

El coeficiente de verosimilitud es la base fundamental del modelo Logit, puesto que presenta la mayor probabilidad entre modelos al utilizar datos de muestra. La diferencia de los cocientes de verosimilitud entre dos modelos se distribuye según el Ji-cuadrado en base a los grados de libertad correspondientes a la diferencia en el número de variables de ambos modelos (Pelaez, 2016).

2.4.2 Modelos de árbol

Los modelos de árbol son precisos, estables y más sencillos de interpretar porque construyen reglas de decisión que se pueden presentar en un diagrama de árbol. La diferencia con el modelo de regresión simple o Logit, es que el modelo de árbol puede representar relaciones no lineales para resolver problemas (Sandoval, 2018). El modelo de árbol puede ser por árbol de decisión simple o el promedio del árbol de decisión, en la siguiente figura se muestra un esquema de este modelo:

Figura 4. Modelo de árbol de decisión



Fuente: (Grau, 2020)

El árbol de decisión puede resolver problemas de regresión y de decisión, esto le otorga una ventaja frente al modelo de regresión, la construcción de un árbol depende de dos pasos:

1. Dividir el espacio de predicción en función de la variable x, que son las variables explicativas, posterior al análisis de la variable x se realizan los cortes y divisiones que finalizan en nodos terminales, como se muestra en la figura anterior.
2. Cada observación debe contener la predicción, que es la media de los valores de respuesta observados en ese nodo; cada nodo terminal tiene una media de valores de respuesta.

La debilidad del modelo de árbol radica en que a medida que crece las divisiones, los resultados se vuelven más confusos y complejos de interpretar, esto incrementa el sesgo en

el modelo, por lo que es necesario realizar una poda para obtener un subárbol, para hallar un modelo con un menor error de prueba (Grau, 2020).

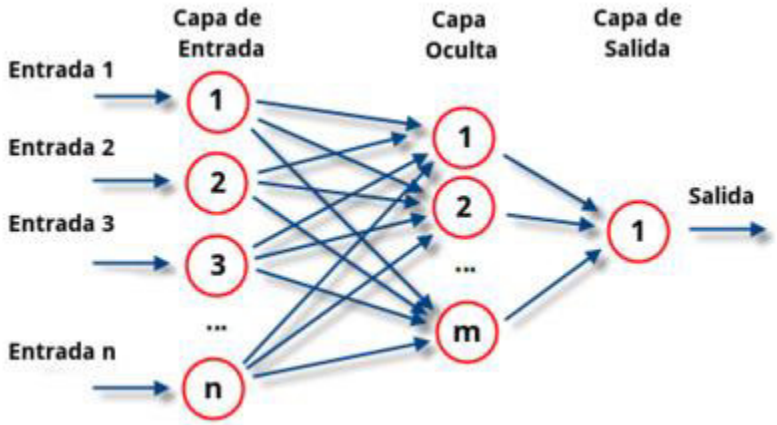
Frente a los modelos lineales, el modelo de árbol presenta ventajas de utilización, ya que su decisión es de mejor interpretación, además la interpretación gráfica permite un mejor análisis frente al modelo lineal. La desventaja radica en su falla al incorporar un alto número de variables o al momento de cambiar de variable, ya que tocaría regresar al punto de partida en el modelo de decisión.

2.4.3 Redes neuronales

El modelo de redes neuronales busca imitar el comportamiento del cerebro, mediante la interconexión en red para enviar mensajes (Sandoval, 2018). A criterio de Aceituno (2019) Este modelo se inspira en la biología, mediante la creación de redes de procesamiento neuronal y la estructura de algoritmos de entrenamiento y recuperación.

La red neuronal se organiza en capas, funciona simulando un elevado número de unidades de procesamiento interconectado, a modo de versiones abstractas de neuronas. Existen tres capas en una red neuronal, la capa de entrada, capas ocultas y capas de salida. La red aprende examinando los registros individuales, generando predicciones para cada registro y realizando ajustes a las ponderaciones cuando realiza una predicción incorrecta. Cuando la red neuronal esta entrenada es capaz de resolver resultados inciertos.

Figura 5. Redes neuronales



Fuente: (Bidaurrazaga, 2019)

Este modelo ha tenido un fuerte crecimiento en su aplicación en medición de riesgo de crédito debido a su habilidad cognitiva para analizar escenarios y predecir resultados basado en el comportamiento de cerebro (Bidaurrazaga, 2019). La desventaja de las redes neuronales radica en su lentitud de aprendizaje y entrenamiento, necesitan una alta capacidad de cómputo para su mejor procesamiento. Con la revolución del Big Data, este modelo ha ganado espacio entre los desarrolladores y programadores.

2.5 Riesgo crediticio

Para encaminar el presente documento es necesario conocer conceptualmente el riesgo crediticio y su implicación en la dinámica de una institución financiera. El eje de la investigación lo constituye el riesgo crediticio junto con la aplicación de machine learning le permita reducir la posibilidad de no pago de los clientes que reciben un préstamo que puede ser destinado para consumo, hipotecario, comercial y/o microcrédito.

El riesgo es innato de cualquier actividad que emprende el ser humano, ya que, se asocia con la incertidumbre y la falta de certeza, según Rivillas y Reina (2016) es entendido como la volatilidad, está presente en todas las decisiones que afrontan las personas y en las organizaciones. Especificando sobre el riesgo de crédito, es la probabilidad de impago en las deudas contraídas con entidades financieras.

El riesgo crediticio parte de un concepto de la gestión integral del riesgo establecido por el Acuerdo de Basilea, esta institución transnacional de supervisión bancaria, indica el aspecto clave en las instituciones financieras es la identificación del riesgo mediante técnicas cualitativas y cuantitativas, esto conlleva, a la minimización de la posibilidad de no pago en una operación crediticia.

Los Acuerdos de Basilea son una serie de directrices que nacen por iniciativa del Grupo de Gobernadores de los Bancos Centrales del G-10 en el año 1974, con el fin de equilibrar y normalizar las políticas de riesgo de las instituciones financieras. Los principales puntos de coincidencia en el Acuerdo de Basilea son medidas de cooperación institucional, adecuación del capital, gestión de riesgos, entre otros aspectos que permiten mejorar la administración bancaria. A lo largo de estos años se han emitido tres directrices en lo que concierne a la gestión de riesgos, estos son Basilea I en el año 1988, Basilea II en el año 2004, Basilea III en el año 2010 y Basilea IV en el año 2018.

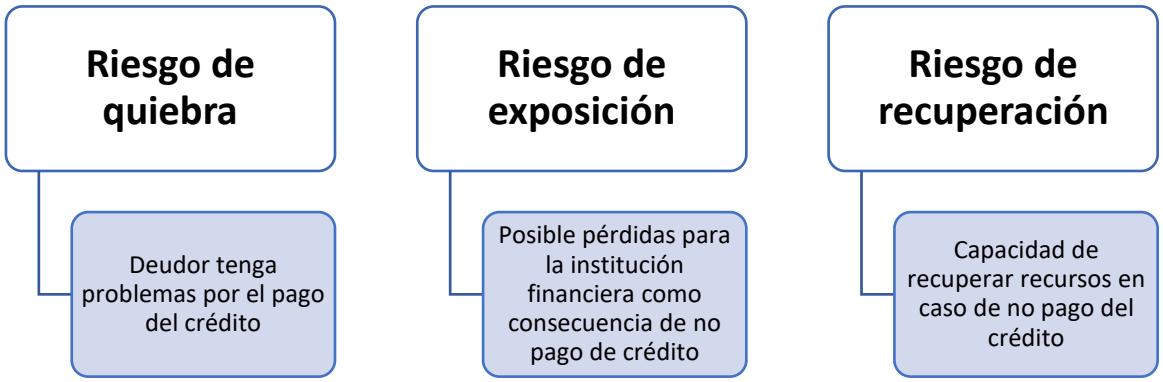
El objetivo principal de los acuerdos de Basilea se orienta en construir un sistema bancario fuerte y equilibrado, que se apegue a normas de crecimiento sostenible, considerando que el sistema bancario es un actor estratégico en el proceso de intermediación financiera, ya que canaliza los recursos de los depositantes hacia las necesidades crediticias (Sánchez, 2018).

Con respecto a la gestión integral de riesgo, Labanda (2017) la define como el proceso de toma de decisiones en base a las expectativas de beneficios futuros, ponderando las posibles pérdidas y evaluando los resultados de forma homogénea y ajustada a la estructura de cada entidad financiera. Asumir la gestión integral de riesgo permite a una entidad financiera mejorar su proceso de toma de decisiones en base de criterios técnicos que se apeguen a certezas y dejen de lado actividades regidas por el azar. En el caso del riesgo crediticio, se busca minimizar la probabilidad de no pago en una operación crediticia.

En el caso ecuatoriano, la Superintendencia de Bancos (2004) establece el riesgo de crédito como posibilidad de pérdida debido al incumplimiento del prestatario en una operación de crédito. La medición del riesgo de crédito va más allá de técnicas contables o financieras, ya que estas técnicas reconocen el riesgo en el momento que sucede un incumplimiento, es por ello, que estas técnicas no son suficientes para evitar la pérdida de valor patrimonial de la institución financiera.

Los factores que se deben considerar para la acertada medición del riesgo de crédito es la probabilidad de incumplimiento, calidad crediticia del deudor, correlación entre incumplimientos, concentración de cartera y tasa de recuperación de deuda (Cisneros, 2016). Como parte de la valoración del riesgo de crédito es necesario asumir los siguientes criterios:

Figura 6. Componentes del riesgo de crédito



Fuente: (Bank for International Settlements, 2013)

El Acuerdo de Basilea establece como preceptos para el control del riesgo de crédito a estos factores:

1. Proporcionar incentivos para la gestión del riesgo y su medición apegada a metodologías cuantitativas y cualitativas.
2. Incrementar la seguridad y solidez del sistema financiero centrándose en el gerenciamiento del riesgo
3. Promover la capitalización de utilidades del sistema financiero
4. Incrementar los procesos de control para la medición del riesgo

2.5.1 Metodologías de valoración del riesgo de crédito

Entre los principales modelos de medición de riesgo, son los conocidos como scoring de crédito, modelos multivariados, arboles de decisión, modelos cuantitativos Probit y Logit y matrices de transición (Rivillas & Reina, 2016).

Para la valoración del riesgo crediticio existen metodologías tradicionales y modernas, en el primer caso se concentran en determinar el riesgo crediticio en base al análisis de variables como capacidad de pago, capital y garantía; asignando puntos a un sistema de calificación conocido como scoring crediticio.

En el caso de las metodologías modernas, se aplican modelos estadísticos que buscan la predicción de resultados, antes de que suceda una perdida asociada al no pago de un crédito, para esta metodología se utilizan variables externas e internas al entorno del solicitante del crédito (Galicia, 2013). En este tipo de metodología es donde los conceptos de machine learning permite mejorar la experiencia y calificación crediticia de un cliente de una entidad financiera.

En ambos casos se utiliza información proporcionada por el cliente e investigación propia de la institución financiera, la gestión documental es clave en la investigación junto con la asunción de información estadística, como probabilidad, desviación estándar y dispersión de datos. Los principios que deben asumir las metodologías de evaluación crediticia se basan en las recomendaciones del Acuerdo de Basilea:

Figura 7. Principios Acuerdo de Basilea

Principios del Acuerdo de Basilea para análisis del riesgo de crédito					
Ambiente apropiado para la gestion de riesgos	Operar un proceso sólido de toma de decisiones	Gestión profesional del riesgo	Agregar valor a la operación de riesgo	Independencia funcional entre areas de riesgo y unidades de negocio	Monitoreo apropiado para el monitoreo y seguimiento

Fuente: (Bank for International Settlements, 2013)

La asunción de estos principios de riesgo crediticio permite la creación de una metodología de evaluación efectiva que contribuya a minimizar el impacto del no pago del crédito por parte de los solicitantes, lo que se busca con las metodologías de evaluación es:

- Determinar el nivel de provisiones
- Mejorar las prácticas crediticias
- Crear segmentos de mercado crediticio en base a la exposición al riesgo
- Establecer parámetros cuantitativos y cualitativos para valorar la exposición al riesgo

La puesta en marcha de las recomendaciones de Basilea junto con los modelos de predicción basados en aprendizaje automático, permiten reducir la probabilidad de que, a su vencimiento, una empresa o una persona no haga frente a un parte o la totalidad de su obligación financiera, debido a reducción de ingresos, posibilidad de quiebra, iliquidez u otra razón asociada a factores intrínsecos del deudor o factores del entorno externo como recesión

o crisis económica (Grau, 2020). En esta consideración radica la importancia de la generación de modelos de predicción con el fin minimizar el impacto de eventos económicos, financieros o sociales en la evaluación crediticia.

2.5.2 Procesos empleados en el otorgamiento de crédito

Para conocer los procesos que se emplean en el otorgamiento de crédito en una institución financiera, se procede a entrevistar a expertos del sector bancario y cooperativo que tienen experiencia y conocen como se desarrolla esta tarea, obteniendo la siguiente información clave.

2.5.2.1 Proceso de Captación Pasiva de Crédito

- En el proceso de captación pasiva intervienen directamente el cliente y la plataforma en la cual debe ingresar con su usuario y contraseña previamente otorgados con el fin de solicitar información acerca de un crédito.

- Una vez ingresado el requerimiento, se entrega la información respectiva al cliente, revisa y analiza la información, el cliente muestra o no interés sobre el crédito, si es positivo se le solicita documentación para iniciar con el proceso, mientras que si es negativo termina el proceso.

- Recibida la documentación por parte del cliente, el primer paso es revisar la central del riesgo del cliente y validar el cumplimiento de políticas de crédito implementadas por la institución.

- Si el cliente no cumple con las políticas o no pasa el filtro de la central de riesgos, se registra en el sistema este hecho y termina el proceso. Si el cliente cumple con las políticas y central de riesgos, se registra y actualiza la base de datos del cliente, y luego de ingresar la solicitud de crédito, se le designa un asesor de negocios para que realice el seguimiento a la solicitud.

- Se imprime la solicitud de crédito y se recepta las firmas del cliente, con el fin de crear el expediente del cliente.

- Con el expediente elaborado, se entrega al asesor designado para que continúe con la evaluación del crédito.

2.5.2.2 Proceso de Captación Activa de Crédito

- El asesor de negocios elabora la hoja de ruta y programa visitas a potenciales clientes, brinda información requerida por el cliente acerca de los créditos.

- Si el cliente está interesado el asesor de negocios solicita documentación para aplicar al crédito y si es negativo el asesor ingresa la información en la hoja de ruta dando por terminado la visita.

- El cliente entrega la documentación al asesor de negocios y continua con el proceso, el primer paso es revisar la central del riesgo del cliente y validar el cumplimiento de políticas de crédito implementadas por la institución.

- Si el cliente cumple con las políticas el asesor deberá llenar la solicitud de crédito respectiva y recoger las firmas del cliente.

- Ingresa la información en la hoja de ruta como la gestión diaria realizada con el cliente y entrega la información a la plataforma.

- En la plataforma se registra y actualiza la información del cliente y se ingresa la solicitud de crédito en el sistema.

- Si el cliente no cumple con las políticas el asesor ingresa en la hoja de ruta con el listado de clientes que no cumple políticas y no se prosigue con el proceso de gestión de créditos.

2.5.2.3 Proceso de evaluación de créditos

- El asesor de negocios elabora la hoja de ruta en el sistema.

- Recopila información del cliente y verifica información cualitativa del cliente para conocer si el cliente es sujeto de crédito, en el caso de que sea así sea se solicita información cuantitativa al cliente y en el caso de que no sea sujeto de crédito se rechaza la solicitud de crédito y se selecciona el motivo de rechazo.

- El cliente entrega documentación cuantitativa solicitada por el asesor de crédito.

- El asesor recibe la documentación y verifica la información para seguir con el proceso, en el caso de que si continúe como sujeto de crédito se llena la solicitud de crédito y se recepta las firmas del cliente.

- El asesor registra en el sistema la evaluación de la solicitud y completa el expediente crediticio, si la evaluación es positiva previo la revisión de informes de crédito y validación de referencias tanto laborales como bancarias, personales y comerciales el

asesor procede a derivar la operación por sistema al Comité de Crédito para su respectiva revisión.

- En el caso de que el asesor registre una evaluación negativa debe notificar al cliente que la solicitud de crédito está rechazada además de ingresar al sistema el motivo del rechazo.

- Ingresa en la hoja de ruta el listado de solicitudes evaluadas de manera positiva y las solicitudes rechazadas para el control.

2.5.2.4 Proceso de aprobación de créditos

- El proceso inicia con el nivel de aprobación en donde se reúne el Comité de Crédito para evaluar y analizar las operaciones derivadas de crédito según su autonomía y aprobación de excepciones.

- Se registra la decisión y resultados de la evaluación que puede ser: aprobado, rechazado y observado.

- Si el crédito es rechazado se ingresa en el sistema para que el asesor de negocios notifique al cliente el motivo del rechazo.

- Si el crédito es observado se registra las observaciones en el sistema para que el asesor regularice lo solicitado por comité e inicie el proceso de aprobación nuevamente.

- Si el crédito es aprobado y no requiere otro nivel de autonomía y excepciones se envía la aprobación vía sistema y se notifica al cliente.

- En el caso de que se requiera otro nivel de autonomía se envía el expediente al analista de riesgo para que revise la solicitud y emita un criterio y envíe con su opinión al comité de crédito y procedan con la revisión.
- Si es necesario el analista de riesgo puede realizar una visita al cliente para comprobar la veracidad de la información y recolectar más documentación o datos que validen su solicitud de crédito.
- El Comité de Crédito se reúne para revisar el expediente que envía el analista de riesgo y procede a aprobar o rechazar la solicitud de crédito.

Los diagramas de flujo de cada proceso en mención se pueden observar en anexos.

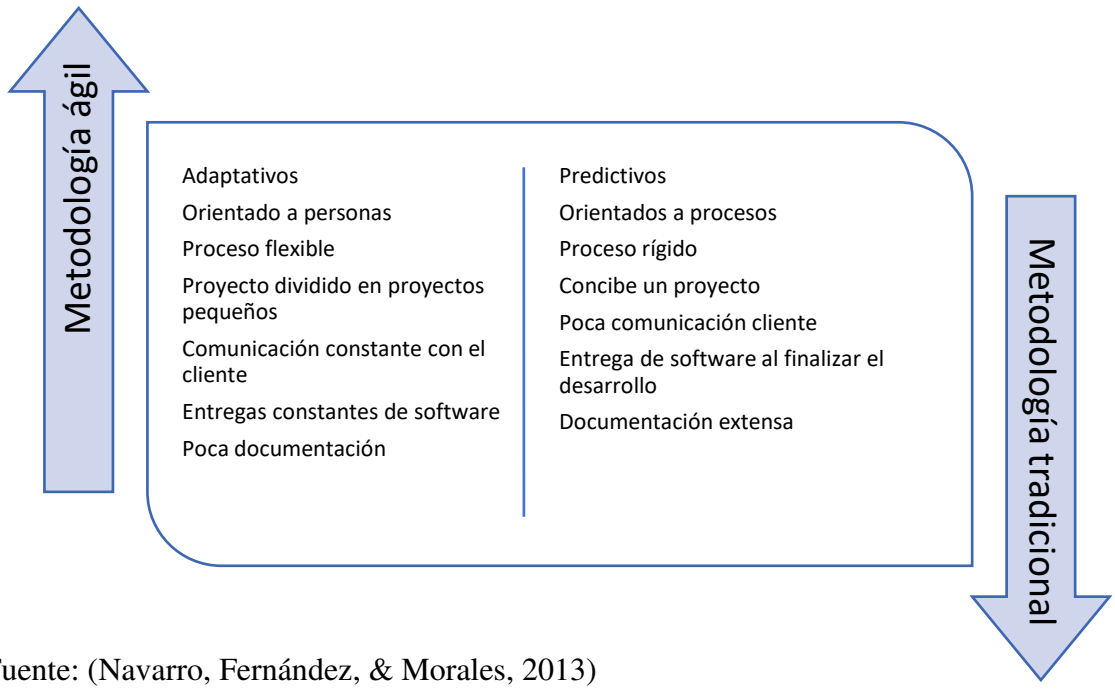
3 METODOLOGIA AGIL

El desarrollo de software ha sido una tarea compleja, que involucra a un grupo multidisciplinario de personas, que se aglutinan alrededor de un objetivo común. En los primeros años de la década de 1990, la metodología de desarrollo de software era una tarea burocrática que se constituían en un gasto innecesario de recursos humanos, materiales y financieros, esto frenaba el interés de las organizaciones por alcanzar una automatización de sus procesos a través de soluciones de ingeniería de software.

Con la puesta en marcha de las metodologías ágiles, se consigue un cambio en la concepción del desarrollo de software como un proceso que se adapta a la necesidad de cliente en base del cumplimiento de tareas enfocadas en la comunicación constante con el cliente. La propuesta de las metodologías ágiles era minimizar el uso de recursos y prevalecer el interés por concluir el proyecto en el menor tiempo posible y con la menor cantidad de recursos, eso

lo diferenciaba sustancialmente de la metodología tradicional. En la siguiente figura se expone las principales diferencias entre ambas metodologías:

Figura 8. Metodología tradicional vs Metodología ágil



Fuente: (Navarro, Fernández, & Morales, 2013)

El documento que compendia el esfuerzo colaborativo de expertos en el desarrollo de software se conoce como el Manifiesto por el Desarrollo Ágil, creado en el año 2001, a partir de esta publicación se reconoce la necesidad de un lineamiento común para alcanzar un mayor suceso en el desarrollo de software. La principal característica de esta metodología se concentra en la cooperación entre el desarrollador y el cliente, la simplicidad en la implementación y en la presentación de los resultados, entregas frecuentes y el desarrollo incremental (Navarro, Fernández, & Morales, 2013).

Las metodologías ágiles con mayor uso en el contexto del desarrollo de software son Scrum y Extreme Programming; los equipos han adoptado estas metodologías, ya sea usadas por

separado, en conjunto o una mezcla de las mejores prácticas de cada una de ellas, en este aspecto precisamente radica la versatilidad de las metodologías ágiles, en su capacidad de adaptación, considerando que Scrum apunta a la administración del proyecto y Extreme Programming está enfocado en el desarrollo.

3.1 Extreme Programming

Extreme Programming conocida por su abreviación XP es una metodología ágil de desarrollo de proyectos que considera la entrega de un proyecto en base de las necesidades del negocio y conforme la calidad en la generación de información; surgió en el año 1996 a partir de las ideas de Kent Beck y Ward Cunningham, que buscaban la implementación de buenas prácticas en la construcción y diseño de software (Laínez Fuentes, 2014).

A partir del año 2007, esta metodología ha sido ampliamente difundida y puesta en práctica por un amplio número de desarrolladores de software. Con el pasar del tiempo, la metodología se ha ido perfeccionando hasta alcanzar un nivel de ejecución que permite el desarrollo de proyectos, fundamentándose en cinco valores, que son comunicación, simplicidad, retroalimentación, respecto y coraje.

La metodología XP tiene como eje central cinco valores retroalimentación, simplicidad, cambio incremental, aceptación del cambio y trabajo de calidad. La práctica de la metodología incluye los siguientes componentes:

- *Planning game:* permite plasmar los alcances de una entrega funcional y el establecimiento de las fechas de cumplimiento, divide las responsabilidades entre los

desarrolladores y el cliente, en base de las historias del usuario, donde se registran los requerimientos del cliente y los desarrolladores se encargan de ejecutarlas. La decisión se adopta mediando la comunicación y la adopción del uso y fuente de recursos necesarios para llevar adelante las necesidades del cliente.

- *Pequeñas entregas:* son los ciclos cortos de desarrollo, donde se muestra el software terminado al cliente, en este punto se implementa un proceso de retroalimentación hasta conseguir las pruebas de aceptación.
- *Diseño simple:* el desarrollo software debe ser simple basado en las historias del usuario.
- *Programación en pareja:* el diseño del software debe ejecutarse entre dos programadores, realizándose una rotación periódica para que el conocimiento fluya entre los participantes.
- *Pruebas:* constituye el componente principal en las historias del usuario donde se interactúa entre el trabajo de los programadores y la necesidad del cliente.
- *Refactoring:* son las acciones correctivas tendientes para realizar cambios que mejoren la estructura del sistema sin afectar el núcleo principal de funcionamiento del software.
- *Integración continua:* cada tarea realizada en el sistema se incorpora a una fase de prueba.
- *Propiedad común:* el código de desarrollo del software pertenece al grupo de desarrolladores, es viable realizar cambios siempre y cuando agreguen valor al sistema.

- *Paso sostenible:* relacionado al ritmo de trabajo del equipo de desarrollo, el canon establecido menciona que no se debe trabajar horas extras más de dos semanas consecutivas.
- *Cliente en sitio:* debe existir un representante del cliente trabajando a tiempo completo con el equipo de desarrollo, coordinando las acciones de las historias del usuario.
- *Metáfora:* el manejo del lenguaje de comunicación por parte de los participantes en el proyecto debe ser unificado.
- *Estándares de código:* normas que establecen la forma a redactar el software, puesto que el código es el punto de partida del proyecto.

La versatilidad de la metodología XP, se basa en la entrega de software cuando el cliente lo necesite y el plazo que los clientes lo requieran, eso permite una agilidad en el desarrollo de software, incluso en fases posteriores del ciclo de vida del proyecto; esta metodología es considerada como parte de las metodologías agiles, que se enfocan en el objetivo, el equipo de trabajo, los recursos disponibles y la participación del cliente.

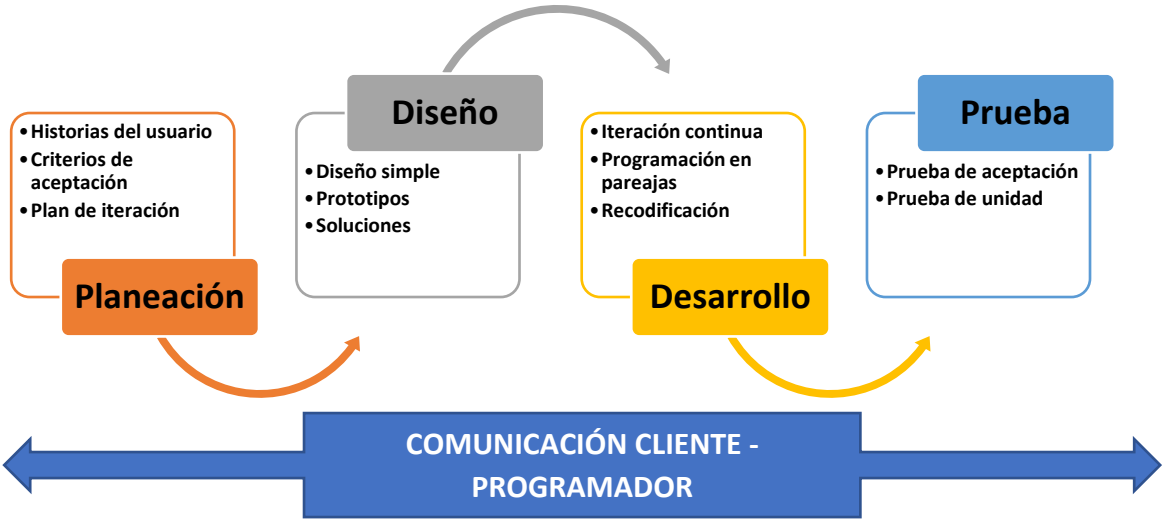
3.1.1 Ciclo de vida del proyecto

El ciclo de vida del proyecto es la etapa donde se busca entender lo que necesita el cliente, valorar el trabajo y plantear la solución (Rodriguez & Miñano, 2017). La metodología XP establece las interacciones con el cliente en las cuatro fases básicas, que son planear (análisis), diseñar, codificar (desarrollo) y prueba; como eje transversal durante todo el proceso se establece a la comunicación entre el cliente y los programadores (Villareal, 2013).

La diferencia principal de la metodología XP con otro tipo de metodología que diseñan el desarrollo del proyecto en base al cumplimiento de etapas, en el caso de XP el proyecto se desarrolla en un ciclo dinámico con bajos intervalos de tiempo denominados iteraciones, en cada iteración se cumple con las fases de análisis, diseño, desarrollo y prueba.

En la siguiente figura se muestra el esquema de funcionamiento de las fases del ciclo de vida del proyecto:

Figura 9. Ciclo de vida del proyecto



Fuente: (Villareal, 2013)

- **Planeación**

En esta fase se organiza la reunión de planning game entre el cliente y el equipo de desarrollo, con la presentación de los requerimientos funcionales o historias del usuario. El equipo de desarrollo debe ordenar las historias del usuario y medir el alcance de las actividades, que se deben ejecutar en la fase de diseño.

Los aspectos claves en la fase de planeación son las historias del usuario, plan de entregas, plan de iteraciones y reuniones diarias de seguimiento (Rodriguez & Miñano, 2017).

- **Diseño**

La iteración inicia con la fase de diseño, donde debe primar la simplicidad de las tareas para la solución de las necesidades del cliente, los miembros del equipo deben aportar ideas para la llevarlas adelante en esta fase, primando la concepción de la simplicidad con el fin de minimizar el impacto de tareas no solicitadas por el cliente o tareas que no aportan soluciones a los problemas encontrados.

- **Desarrollo**

En la fase de desarrollo se implementan códigos para designar el producto funcional, la programación por pares permite la minimización de errores; el papel del cliente es primordial, porque debe estar disponible durante todo el proyecto, para generar un proceso de comunicación constante.

Los aspectos claves en esta etapa son el uso de estándares que facilitan la tarea de los programadores, ejecución de tareas en pares y programación dirigida hacia las pruebas para entender cuáles son los estándares del test desde el principio.

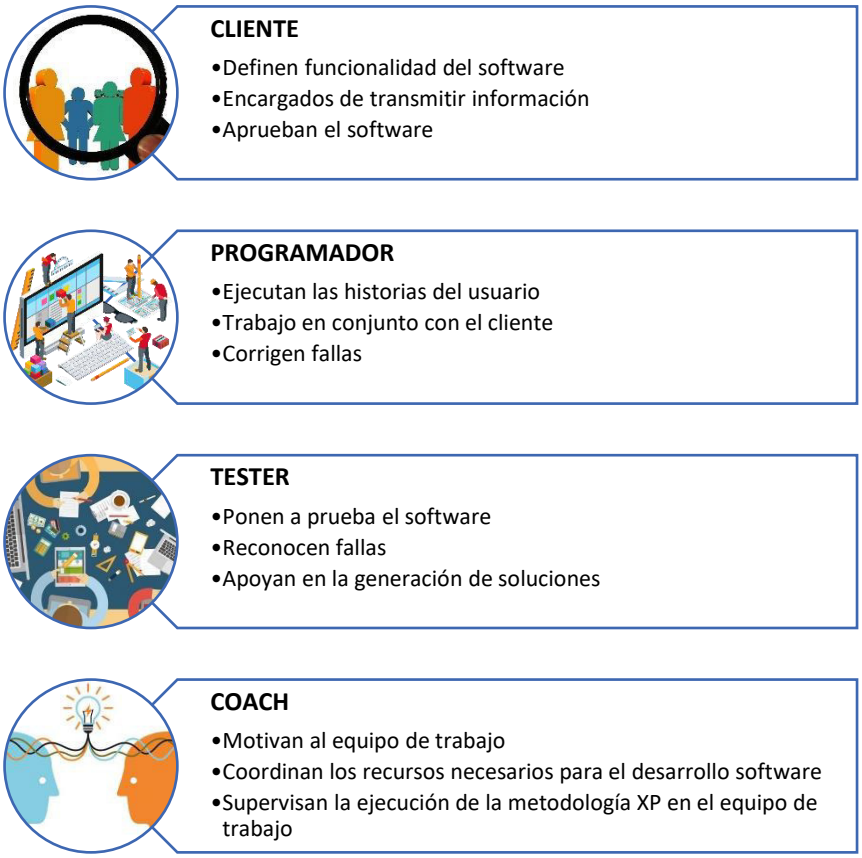
- **Prueba**

Todo el desarrollo del proyecto debe pasar por pruebas unitarias antes de ser testeado por el cliente y previo a la liberación del código. El cliente es el responsable de validar las pruebas. Las historias del usuario creadas en la fase de planeación son cerradas con el cliente tiene el producto a cabalidad.

3.1.2 Roles en XP

Las funcionalidades de cada miembro del equipo de trabajo en el desarrollo de software bajo XP cumplen una tarea específica, que se presenta en la siguiente figura:

Figura 10 .Roles en XP



Fuente: (Villareal, 2013)

Todos los roles dentro de la metodología XP forman un solo equipo de trabajo, es indispensable un representante del cliente y la participación de expertos que conocen el giro del negocio, se debe designar la persona encargada del juego de la planificación (planning game) donde inicia todo el proceso de desarrollo del software.

3.2 CRISP

La minería de datos es un campo de la ciencia estadística y computacional, que permite explorar y descubrir patrones en grandes cantidades de datos, utiliza técnicas de aprendizaje automático, inteligencia artificial y sistema de bases de datos (Gallardo Arancibia, 2014). La sistematización del proceso en la minería de datos es un punto estratégico en la planificación

y ejecución de proyectos, ya que permite organizar la información y extraer datos ocultos o implícitos que a simple vista o por métodos estadísticos tradicionales pasarían inadvertidos (Moine, 2014).

Entre las metodologías para ejecutar proyectos asociados con minería de datos se encuentra Cross Industry Standard Process for Data Mining (CRISP-DM), es una guía de referencia que fue implementada por un grupo de empresas europeas que se reunieron en el año 2000, para generar un patrón en el uso de información y exploración de datos; como metodología describe fases normales de un proyecto y como modelo de proceso determina un ciclo vital de minería de datos (Rodriguez & Miñano, 2017).

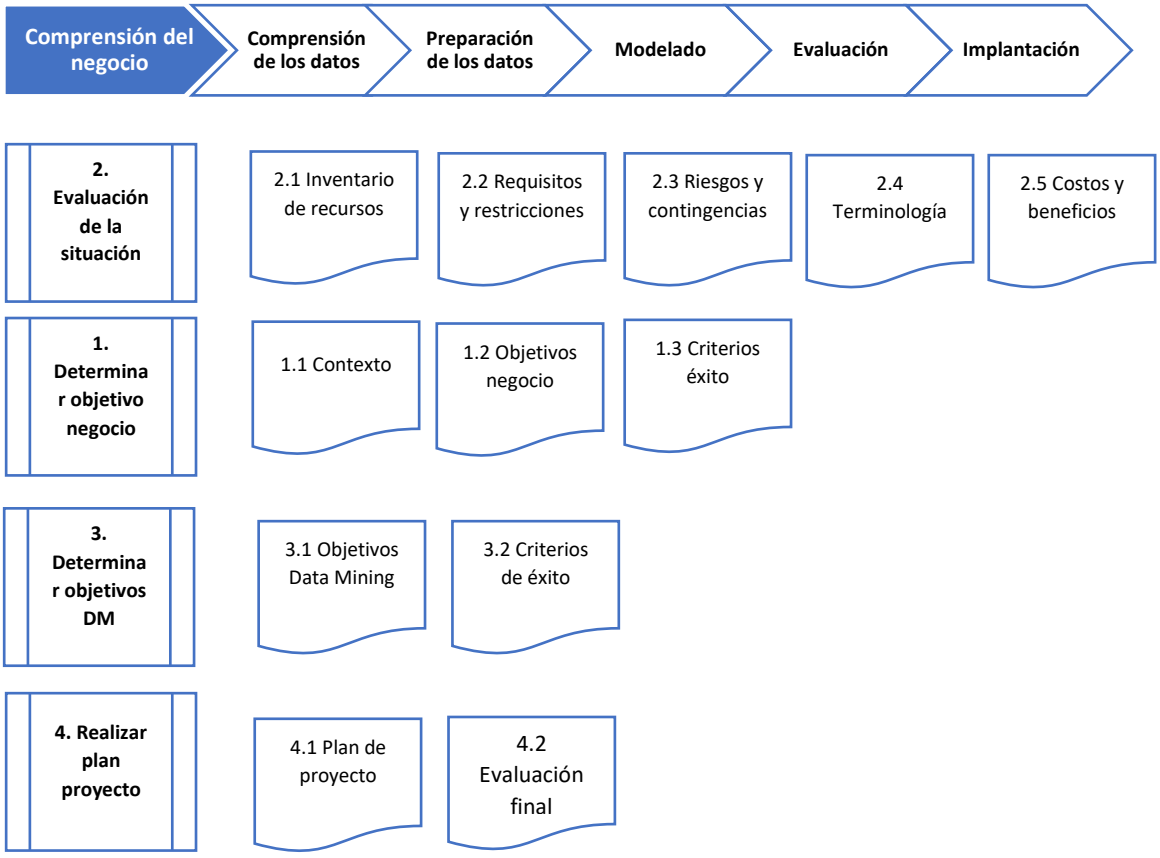
3.2.1 Fases de la metodología CRISP

a. Comprensión del negocio

El objetivo de esta fase es conocer los factores que pueden influir en el resultado del proyecto, en esta fase se formula los objetivos y requerimientos desde una perspectiva del negocio (Rodriguez & Miñano, 2017).

Esta fase es la más importante, ya que permite convertir los objetivos del negocio en acciones técnicas para plantear el plan del proyecto, el aspecto central de esta fase es conocer el problema que se quiere resolver (Gallardo Arancibia, 2014). Las actividades que se deben cumplir en esta fase son las siguientes:

Figura 11. Comprensión del negocio



Fuente: (Gallardo Arancibia, 2014)

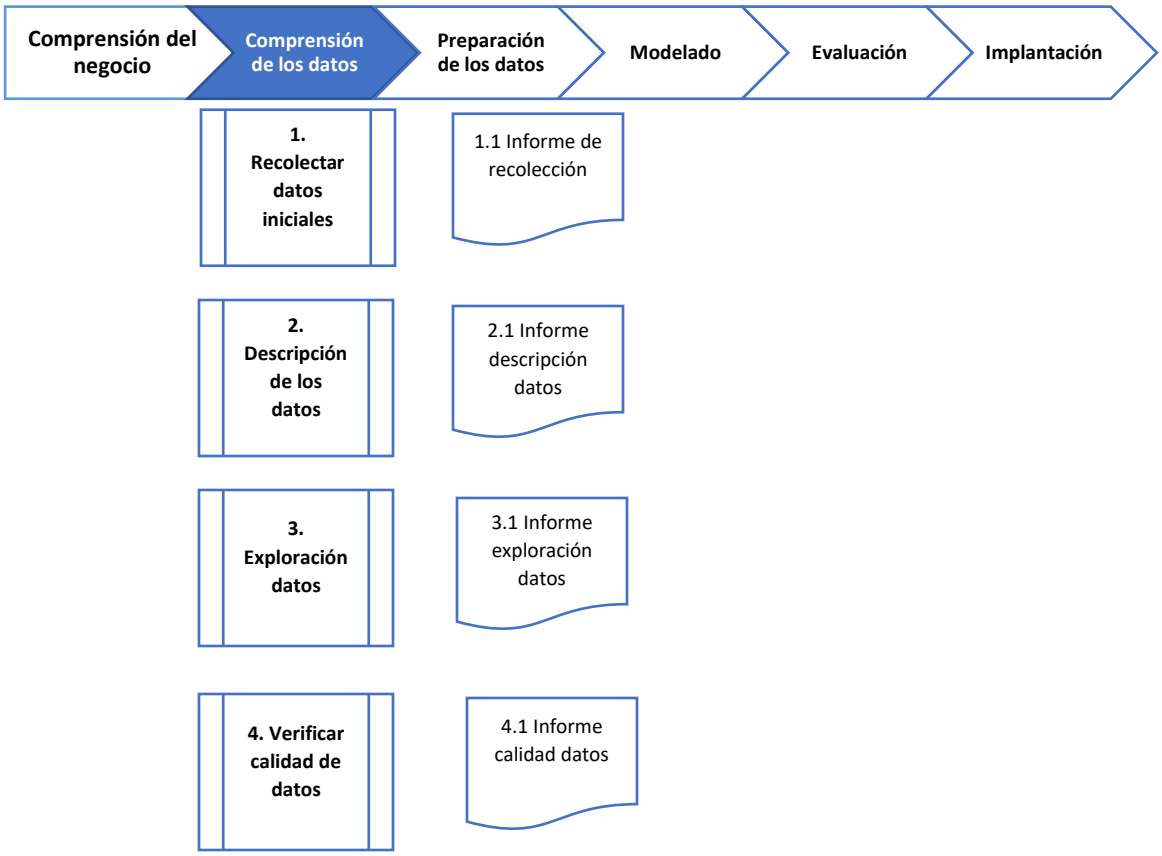
1. Determinar objetivos del negocio: conocer cuál es el problema que se deber resolver, cuál es la necesidad de utilizar la minería de datos, entender el modelo de negocio del cliente para ofrecer una solución final.
2. Evaluación de la situación: calificar el estado de la situación actual antes de la puesta en marcha del proyecto, evaluar los recursos disponibles, la relación costo beneficio entre la situación actual y la propuesta.
3. Determinación de los objetivos: plantear los objetivos en términos de metas alcanzables y en función de la situación que se pretende mejorar en el cliente.

4. Realizar plan del proyecto: elaborar el documento donde conste los pasos a seguir y las técnicas empleadas en cada actividad.

b. Comprensión de los datos

Esta fase enfoca los esfuerzos del grupo de trabajo por acceder a los datos y explorarlos en base al uso de tablas y gráficas, con el fin de evaluar su calidad y viabilidad para el proyecto.

Figura 12. Comprensión de los datos



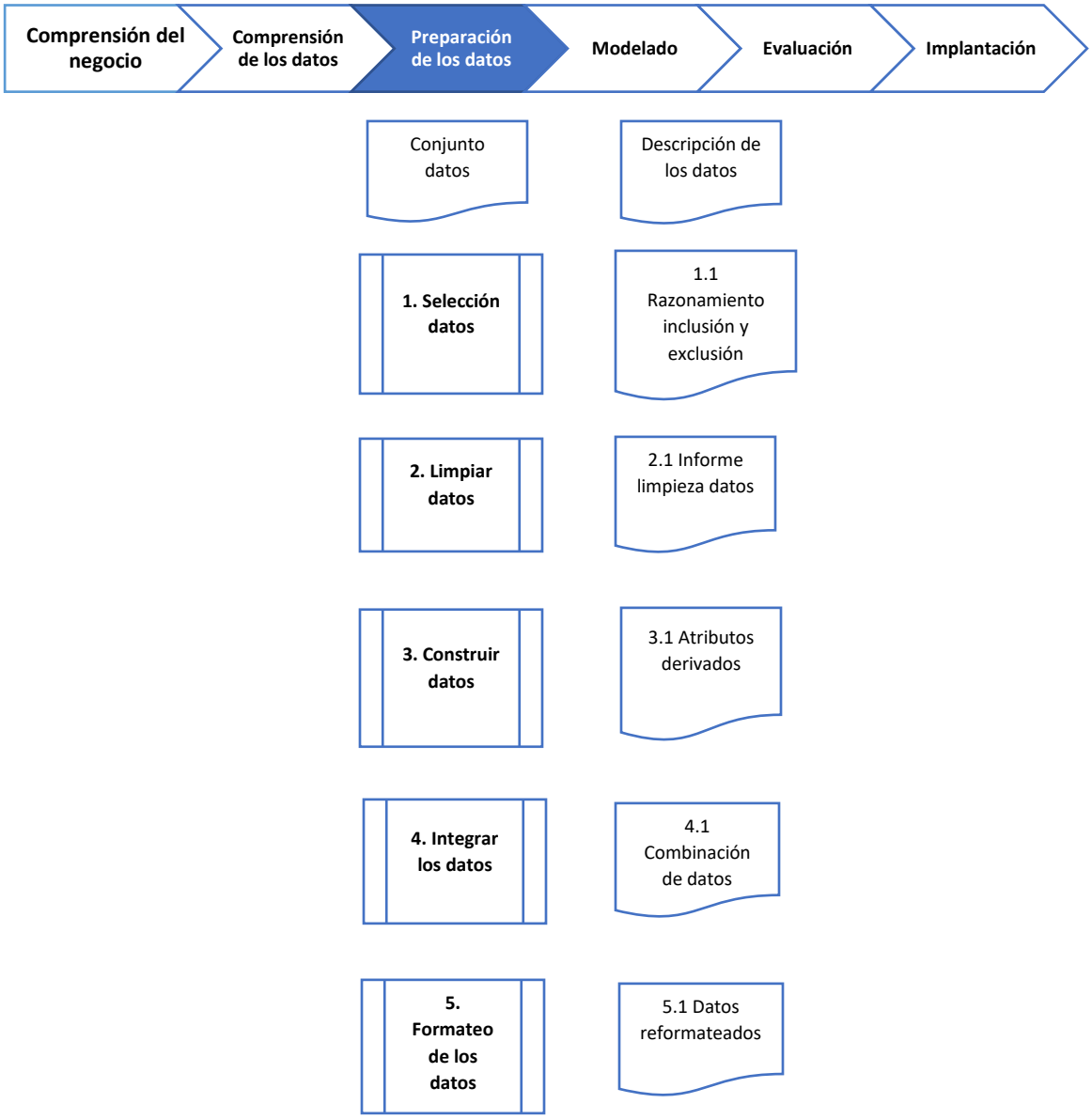
Fuente: (Gallardo Arancibia, 2014)

1. Recolectar datos iniciales: consiste en ubicar los datos necesarios para llevar adelante el proyecto, su localización, técnicas de recolección y detallar problemas que se pueden presentar y están asociados a su utilización.
2. Descripción de los datos: este proceso involucra medir el número de registros y campos de registros, identificación, significado de cada campo y la descripción en formato inicial.
3. Exploración de los datos: permite encontrar una estructura general de los datos con la aplicación técnicas estadísticas básicas.
4. Verificar la calidad de los datos: medir la consistencia de los datos antes del desarrollo del modelo, con el fin de evitar resultados adversos en el informe final.

c. Preparación de los datos

La fase de preparación de los datos utiliza la mayor cantidad de tiempo y esfuerzo en el proyecto, se estima que toma entre el 50% y 70% del total empleado (Rodriguez & Miñano, 2017). La preparación de datos incorpora las actividades de limpieza de datos, clasificación de datos, generación de variables, integración y cambio de formato. Las técnicas utilizadas incluyen visualización de datos, búsqueda de relaciones entre variables y medidas de exploración (Gallardo Arancibia, 2014).

Figura 13. Preparación de los datos



Fuente: (Gallardo Arancibia, 2014)

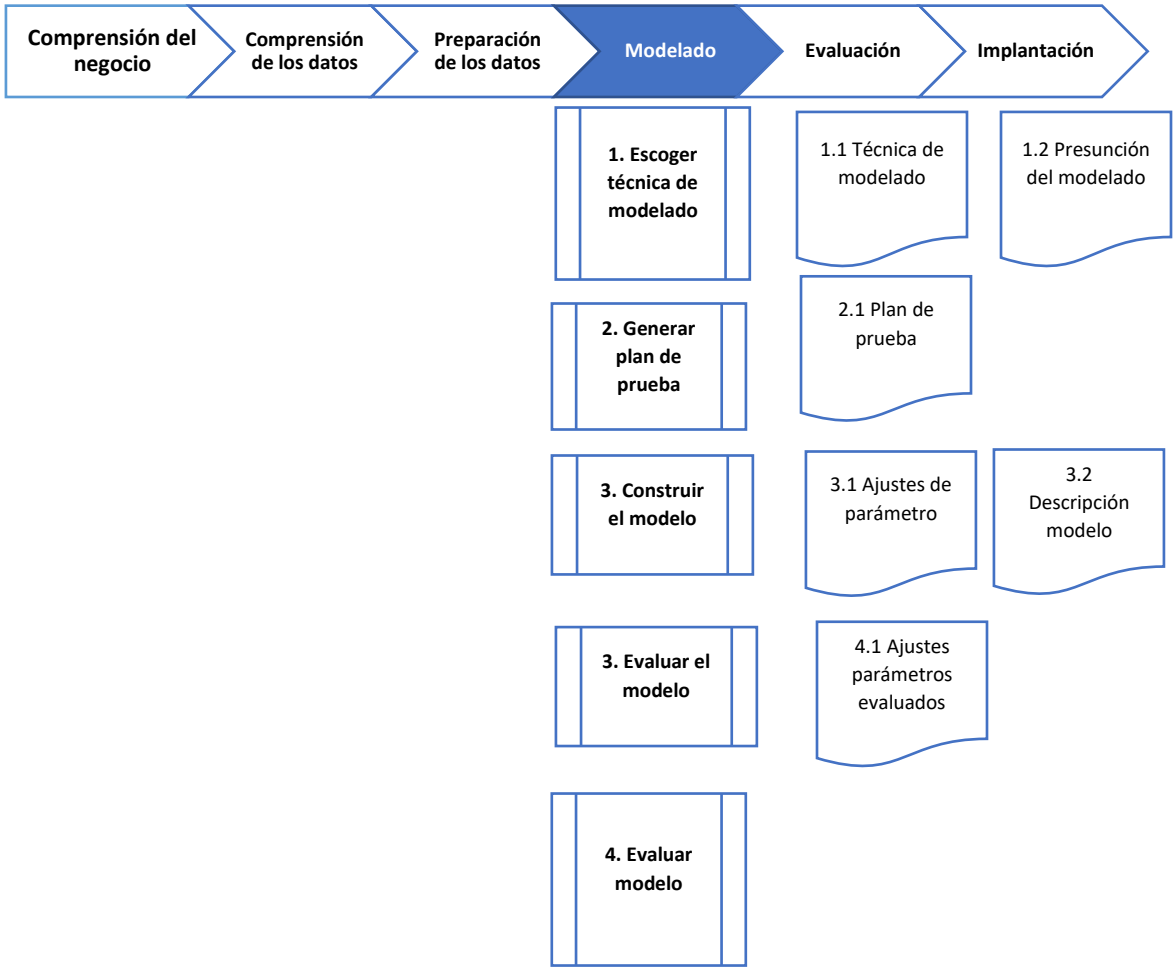
1. Selección de datos: seleccionar del grupo de datos obtenido en la etapa previa los datos a trabajar, basándose en puntos de vista de calidad, corrección y limitaciones en el uso de datos.

2. Limpieza de datos: es una tarea complementaria a la anterior y consume una alta cantidad de tiempo y esfuerzo, incluye tareas de normalización, discretización, reducción de volumen, entre otros.
3. Construcción de nuevos datos: generar nuevos atributos en base de la limpieza de datos, integración de registros y transformación de valores.
4. Integración de datos: en base de la fusión y adición de datos con el fin de unir en conjuntos a los datos con atributos similares.
5. Formateo de los datos: es el paso final antes de la construcción del modelo, es útil para comprobar las técnicas que se requiere aplicar en el modelo, consiste en la realización de transformaciones sintácticas sin modificar el significado de los datos.

Modelado

Los datos preparados en las fases anteriores se incorporan a las herramientas analíticas, en busca de resultados que permitan la solución del problema planteado en el la fase inicial (Rodriguez & Miñano, 2017), las técnicas utilizadas deben cumplir con los criterios de ser apropiadas para el problema que se pretende solucionar, disponer de datos adecuados, cumplir los requisitos del problema, tiempo adecuado para solucionar el problema y conocimiento de la técnica (Gallardo Arancibia, 2014).

Figura 14. Modelado



Fuente: (Gallardo Arancibia, 2014)

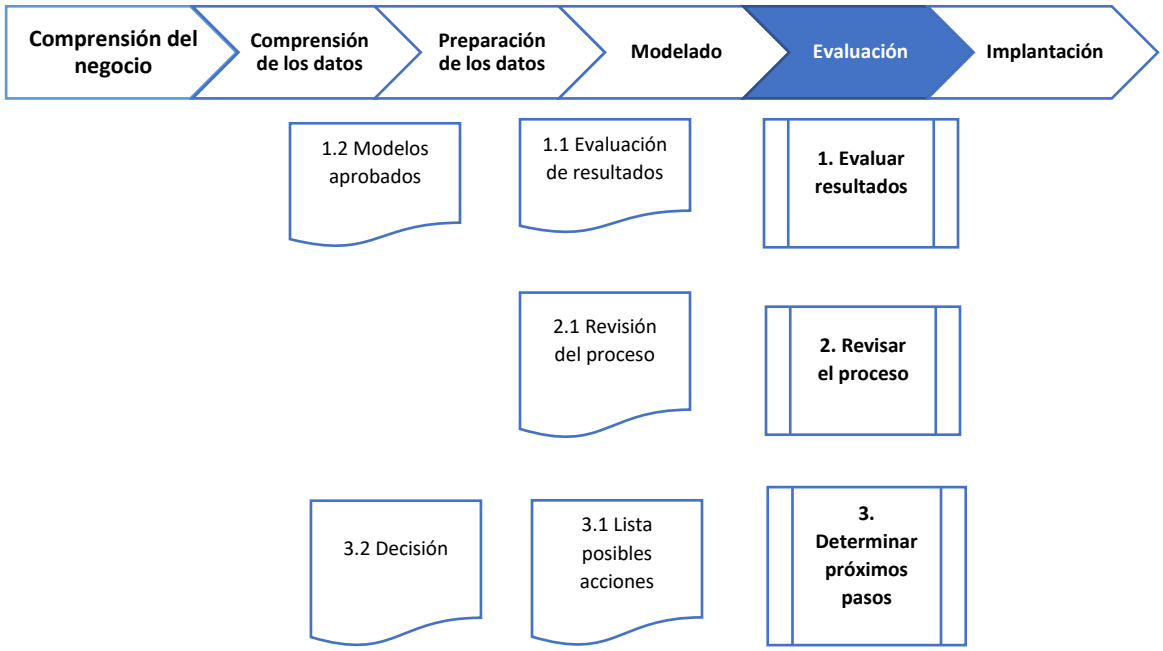
1. Selección de técnicas: la selección de la técnica tiene relación con el problema que se pretende resolver, el objetivo trazado y las herramientas empleadas en la minería de datos; si el problema es clasificación se puede elegir árbol de decisión, si el problema es predicción la técnica puede ser análisis de regresión, si el problema es segmentación puede utilizarse redes neuronales o técnicas de visualización.
2. Generación plan de prueba: es una guía para probar la calidad y validez del modelo, existen dos criterios para probar la validez, los criterios en base de la bondad y en base de la definición del modelo.

3. Construir el modelo: se ejecuta sobre los datos previamente preparados, el modelo o los modelos creados deben basarse en un conjunto de parámetros y descripción de los resultados.
4. Evaluar el modelo: trata sobre la creación de un método que permita la valoración, basado en criterios generados en el plan de pruebas.

e. Evaluación

Corresponde a la verificación de los resultados obtenidos en el modelo sean viables para el cliente y que solucionen el problema planteado en la fase inicial.

Figura 15. Evaluación



Fuente: (Gallardo Arancibia, 2014)

1. Evaluación de los resultados: verificar si los resultados obtenidos cumplen con la solución planteada en función de la necesidad del negocio, para cumplir esta tarea se debe remitir a los objetivos y a los parámetros iniciales, con los cuales se plantea la solución en la fase de comprensión del negocio.

2. Proceso de revisión: identificar elementos que se pueden mejorar en función de pruebas previas que plantea la metodología CRISP en casos de estudio anteriores.
3. Determinar los próximos pasos: en este punto se puede tomar dos caminos, el primero es incorporar los resultados al proceso comercial del cliente o refinar los criterios del modelo en el caso de que no se cumpla con los objetivos planteados.

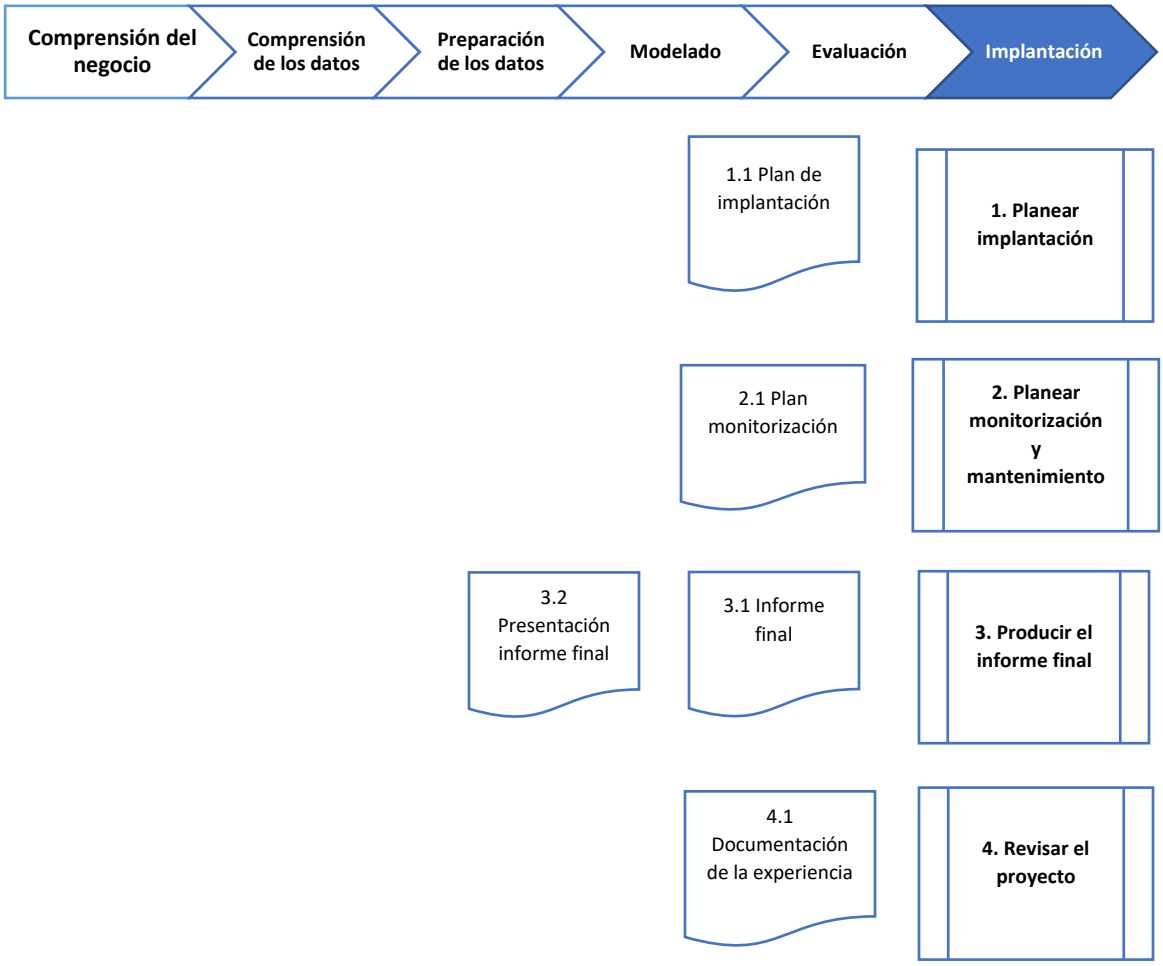
f. Implantación

Esta fase describe las actividades que permiten transformar el conocimiento del modelo en acciones concretas dentro del modelo de negocio del cliente o se ejecuten cambios en el proceso organizativo. Las dos actividades principales en la fase de implantación es la planificación y control de los resultados y la finalización de tareas con la presentación de un informe final y la posterior entrega al cliente.

1. Planificación de la implantación: la correcta implantación incluye las actividades de resumir los resultados, planificar paso a paso la distribución e integración con sus sistemas, difundir los resultados a personas clave en la organización, documentar los resultados e identifique los posibles planes de contingencia.
2. Planificación de monitorización y del mantenimiento: registrar elementos que permiten medir y controlar la validez del modelo, planes asociados a la actualización o cambio en la gestión de información de los datos y tiempo de vigencia del modelo.
3. Creación del informe final: la estructura del informe final debe incluir la descripción detallada del problema original, procedimiento utilizado en el modelo de minería de datos, costos asociados al proyecto, comentarios y resumen de los resultados junto los riesgos y contingencias asociados a la implementación del modelo.

4. Revisión del proyecto final: es el paso final de la metodología e incluye las impresiones y comentarios de los participantes en el modelo de minería de datos.

Figura 16. Implantación



Fuente: (Gallardo Arancibia, 2014)

Respecto a la valuación del modelo de predicción de riesgo de crédito con base al uso de algoritmos de machine learning, la literatura muestra un gran número de modelos aplicables al análisis de riesgos crediticio (i.e. Modelo de Altman, Modelo Logit, modelo de credit scoring, etc.) de los que podemos encontrar sus aplicaciones bien documentadas en la literatura científica ((E. Altman, R. Haldeman y P. Narayanan, 1977), (Alfred DeMaris,

1992), (Chun-Ling Chuang y Rong-Ho Lin, 2009)). Dado que el modelo de credit scoring toma en consideración la conceptualización y calificación de características numéricas y no numéricas que faciliten el enriquecimiento de datos para medir el riesgo crediticio del solicitante, y que, además, se dispone de un conjunto de datos históricos se ha considerado que este modelo (credit scoring) se ajusta a los requerimientos de esta investigación.

A continuación, se describen tópicos relevantes que nos permitirán identificar las variables que se utilizarán para el desarrollo del aplicativo propuesto.

3.3- Recolección de datos

Para realizar el análisis cualitativo es necesario recolectar información que será de utilizada para orientar el desarrollo de la propuesta final. En el presente trabajo de investigación, se utilizó:

- Observación
- Encuestas

3.3.1- Observación

Esta técnica consiste en la observación del fenómeno o hecho para registrar la información y analizarla; permite recabar un mayor número de datos respecto a las variables de estudio. Para este trabajo, se busca observar el procedimiento utilizado para la recolección de información de los clientes previa al análisis de riesgo crediticio. Esto nos permitirá revisar los diferentes factores que los expertos toman en cuenta y que se utilizan en el análisis respectivo.

Para esta investigación se ocupó la observación directa del proceso de análisis crediticio de los clientes. En el siguiente tabal se pueden revisar los detalles más relevantes de este proceso:

Tabla 2: Resultados del proceso de observación directa

Proceso observado	**Resultados obtenidos**
Revisión de información	Nos ha permitido conocer los pormenores que son considerados en el modelo de realizar la revisión de la información otorgada por el cliente. Concretamente nos ha resultado de utilidad la forma como se recolecta la información y la importancia que se le da a cada ítem de información.
Elementos del análisis de crediticio	Principalmente nos ha resultado de utilidad varios elementos revisados en el análisis como lo son: • Estados de cuenta • Comprobantes de ingresos • Comportantes de garantías
Variables de interés	La empresa recoleta información de un gran número de variables, las cuales han sido dividas en tres grandes grupos: • Personal • Crediticio • Historial
Data histórica	Un aspecto muy relevante ha sido revisar la cantidad de información (observaciones) de las que dispone la empresa, y

	se pudo constatar que el número de registros del dataset se ubica en unos 840 mil.

3.3.2- Encuestas

Se ha utilizado, como instrumento de recolección de datos la encuesta, el cual es un documento prediseñado en el que se detalla los aspectos más relevantes a recolectar y que serán de gran utilidad para el desarrollo de la propuesta. En este caso, se recolecta información relacionada con las diferentes variables relevantes para el diseño del modelo predictivo. El instrumento utilizado se puede ***observar en los anexos***.

Se han aplicado un total de 12 encuestas, de las cuales 3 corresponden al personal de la empresa y el resto (9) corresponden al personal de dos empresas amigas del ramo, que accedieron a colaborar en la investigación.

Los resultados de las encuestas se han separado en 5 grupos, como se indica a continuación:

- Tipos de datos
- Variables relacionadas con lo personal
- Variables relacionadas con el historial creditico (grupo 1 y 2)
- Herramientas usadas por el grupo de encuestados

Los datos tabulados se pueden revisar en las figuras desde la 17 a la 21.

Se observa que, a las variables del conjunto de datos históricos, el grupo de focalización, le ha dado un 50% de importancia, seguida de los datos personales, quedando en tercer lugar la data crediticia histórica (ver figura 17).

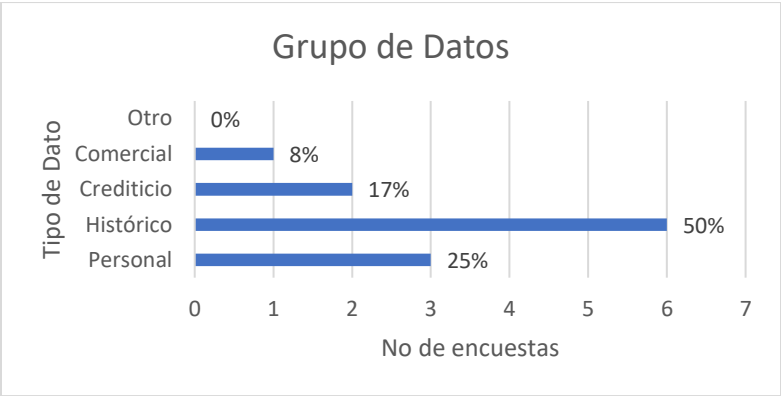


Figura 17: Resultados de la encuesta para el tipo de datos

En cuanto a datos personales, se puede mencionar que las tres variables que sobresalen corresponden a:

- Ingresos mensuales
- Ingresos adicionales
- Relación gastos-ingresos

Esto se puede verificar en la figura 18..

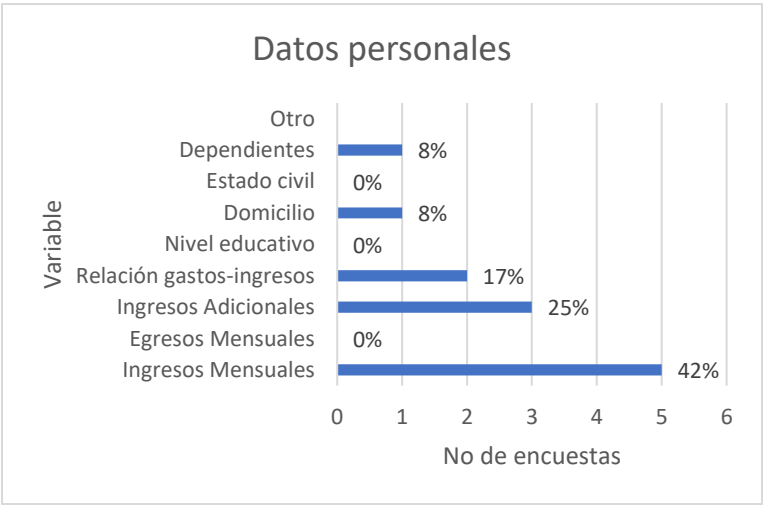


Figura 18: Resultados de la encuesta para las variables personales

En el caso del primer grupo de variables históricas, se aprecia el mayor impacto sobre las variables:

- Créditos obtenidos
- Calificación del buro
- Morosidad
- Cupo de la tarjeta de crédito

Esto se puede verificar en la figura 19.

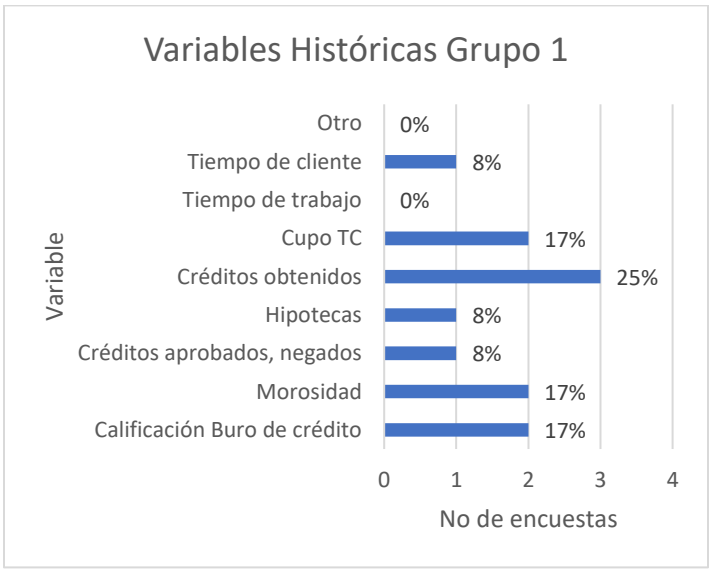


Figura 19: Resultados de la encuesta para las variables históricas grupo 1

En el segundo grupo, el porcentaje de importancia lo acaparan las viables ***Monto del crédito*** y ***Tasa de interés***, con más del 58% para las dos variables (ver figura 20).

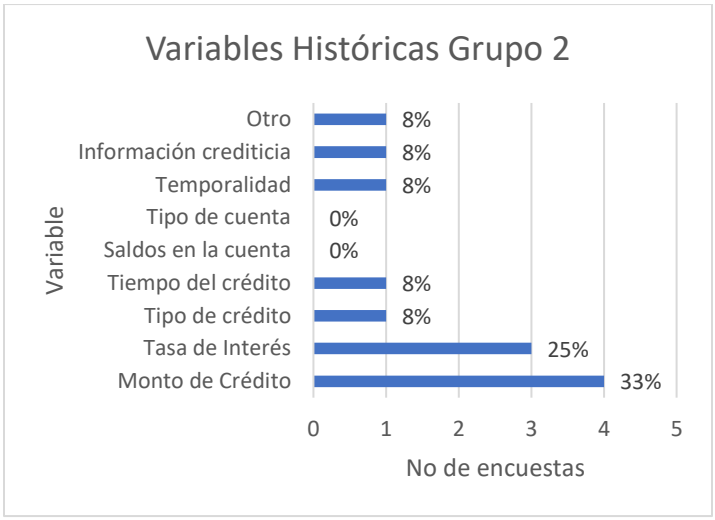


Figura 20: Resultados de la encuesta para las variables históricas grupo 2

Finalmente hay que mencionar que, para este caso particular, el uso de herramientas computacionales parece ser utilizadas por más del 50% de los integrantes del grupo focalizado que ha sido ocupado en esta investigación (ver figura 21).

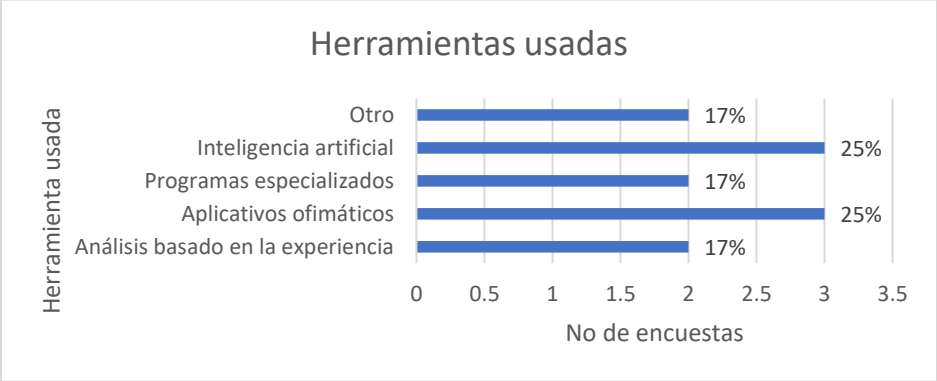


Figura 21: Resultados de la encuesta para los tipos de herramientas usadas

4-. DESARROLLO DEL MODELO DE PREDICCIÓN DE RIESGO DE CRÉDITO

En este capítulo se describen las acciones que han sido tomadas para el desarrollo del modelo de aprendizaje automático que ha resultado del análisis previo, para ello se desglosan cada una de las fases de la metodología CRISP que se ha empleado en esta investigación.

Se debe aclarar que en el capítulo 3 ya ha sido detalladas las fases:

1. Comprensión del negocio
2. Comprensión de los datos

Por lo tanto, aquí nos ocuparemos de las fases restantes.

1. Preparación de los datos
2. Modelado
3. Evaluación
4. Implantación

4.1 Preparación de los datos

La base de datos que será utilizada proviene de la empresa Redemax Cia. Ltda. y consta de 838.860 observaciones con 18 variables para préstamos emitidos entre 2010 y 2021. Se encontró que 61176 clientes del total entraron en mora o se registraron importantes retrasos en el pago, en el caso de algunos de ellos incluso más de 120 días.

De este conjunto de datos se puede decir que se tiene un total de 14 variables numéricas y 3 categóricas. De las variables de tipo numéricas, encontramos las que describen la evolución anual ingresos, comportamiento de pago, como el número de meses desde el último pago morosidad, el número de líneas de crédito abiertas, el número de registros públicos; es decir, Buró de Crédito, así como características del producto como la tasa de interés o el monto del

préstamo. Entre las variables categóricas consideradas están la antigüedad en el empleo, la propiedad o el grado de riesgo expresado por la puntuación FICO (Score, 2022).

4.1.1 Variables utilizadas en el modelo de riesgo crediticio

Para alimentar el modelo de riesgo crediticio, se requiere el uso de variables que corresponde a la información del cliente. Esta información puede ser obtenida por dos vías: la primera corresponde a la información que entrega el cliente en su solicitud de crédito y la segunda forma de obtener la información es a través de la base de datos de la institución financiera. La captura de datos de las variables es un proceso clave, ya que permite alimentar al modelo de información viable y actualizada.

Las variables de carácter financiero son parte fundamental para el desarrollo de un modelo de riesgo crediticio (Rodríguez M., Piñeiro Carlos y De Llano P., 2014). Para determinar las variables más relevantes que intervienen en el modelo de evaluación de riesgo de crédito se ha trabajado en forma conjunta con el área de riesgos identificando los siguientes campos como se muestra en la Tabla 3.

Tabla 3. Variables para el modelo de riesgo crediticio

Número de variable	Variable	Grupo variable
1	Ingresos Mensuales	Personal
2	Ingresos Adicionales	Personal
3	Relación gastos-ingresos	Personal
4	Monto de Crédito	Crédito
5	Tasa de Interés	Crédito
6	Calificación Buro de crédito	Historial
7	Última calificación Buro de crédito	Historial
8	Ultimo monto crédito	Historial

Número de variable	Variable	Grupo variable
9	Morosidad 6 meses	Historial
10	Morosidad últimos 2 años	Historial
11	Total, cargos por mora	Historial
12	Solicitudes negadas	Historial
13	Número de créditos en la empresa	Historial
14	Hipotecas actuales	Historial
15	Número total de créditos	Historial
16	Cupo rotativo TC	Historial
17	Porcentaje de uso de cupo TC	Historial

Las variables seleccionadas, que han resultado de la revisión del proceso aplicado por el departamento de créditos de la empresa, se han identificado 17 variables empíricas (véase la **Tabla *3***). Estas variables, están divididas en tres grupos, que corresponde a la fuente y uso de la información. Las variables denominadas crédito son asociadas directamente con el préstamo solicitado en la institución financiera; el grupo de variables denominadas personal corresponde a información propia del solicitante, esta información es receptada en la solicitud y las variables historial corresponde a la información del cliente disponible en los buros de crédito. La selección de estas variables (a parte de la pericia del personal encargado de evaluar los créditos), han resultado del análisis obtenido en la aplicación de las técnicas de recolección de información descrita en la sección anterior por medio de la observación, y encuestas. Desde el punto de vistas de la literatura científica, hemos encontrado alguna similitud con las variables utilizadas por otros autores que han abordado este mismo problema ((Nisha Arora y Pankaj Deep Kaur, 2020), (Hongmei Chena y Yaoxin Xiang, 2017)).

Por otra parte, desde el punto de vista de la importancia de estas variables para la construcción del modelo (i.e. al aplicar un análisis de importancia de predictores). Considerando el uso de un árbol de decisión para determinar la importancia de una determinada variable o característica, el ranking de selección se calcula como la disminución de la impureza del nodo ponderada por la probabilidad de alcanzar ese nodo. La probabilidad del nodo se puede calcular por el número de muestras que llegan al nodo, dividido por el número total de muestras. Cuanto mayor sea el valor, más importante será la variable. Esta es una técnica muy utilizada que permite determinar el nivel de importancia de cada variable usada en el modelo ((B. Venkatesh y J. Anuradha , 2019), (Nooritawati Md Tahir, Aini Hussain, Salina Abdul Samad, Khairul Anuar Ishak y Rosmawati Abdul Halim, 2006)).

Está claro que estos 17 predictores darían un aporte significativo al modelo, como podemos verlo en la **Figura 22**.

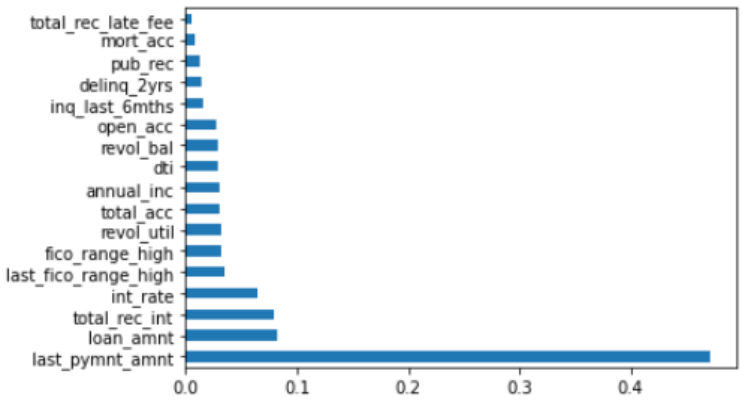


Figura 22: Aporte de cada variable al modelo

Se aprecia que la variable que registra el mayor aporte al modelo es ***Ultimo monto crédito*** (last_pymnt_amnt).

La descripción de las variables descritas en la tabla 2 es la siguiente:

1. Ingresos anuales: cantidad de dinero que recibe anualmente que está relacionado con su ocupación.

2. Ingresos adicionales: cantidad de dinero que recibe anualmente que está relacionado con otras actividades como arriendos, regalías o inversiones.

3. Relación gastos - ingresos: Una relación calculada utilizando los pagos mensuales totales de la deuda del prestatario sobre las obligaciones totales de la deuda, excluyendo la hipoteca y el préstamo LC solicitado, dividido por los ingresos mensuales auto informados del prestatario.

4. Monto crédito: cantidad de dinero que demanda el cliente a la institución financiera, según la información que llena en la solicitud de crédito.

5. Tasa de interés: según la tabla especificada por la institución financiera de acuerdo con el tipo de crédito, puede ser microcrédito, consumo, hipotecario, comercial o educativo.

6. Calificación en el Buró de crédito: score que otorga el Buró de Información Crediticia se calcula mediante varios parámetros que miden de manera sistemática la posibilidad de no pago de un cliente en un rango de 1 a 999.

7. Última calificación Buro de crédito: calificación anterior obtenida en el Buró de Información Crediticia previa a la calificación con la que el solicitante realiza la solicitud de crédito.

8. Ultimo monto crédito: Monto total de pago realizado en el último préstamo.

9. Morosidad 6 meses: Número de veces que el solicitante a incurrido en morosidad en los últimos seis meses.

10. Morosidad últimos 2 años: Número de veces que el solicitante ha incurrido en morosidad en los últimos dos años.

11. Total, cargos por mora: Valor cancelado por indemnización que se le impuso por los daños producidos por el incumplimiento de obligaciones dinerarias.

12. Solicitudes negadas: Número de veces que se le negó al solicitante un crédito en la entidad.

13. Número de créditos en la empresa: El número de líneas de crédito abiertas en la institución en el archivo de crédito del prestatario.

14. Hipotecas actuales: Número de hipotecas que registra el prestatario en las diferentes instituciones financieras del país.

15. Número Total de créditos: El número de líneas de crédito abiertas en cualquier institución en el archivo de crédito del prestatario.

16. Cupo rotativo TC: Monto total que tiene disponible en su tarjeta de crédito y es otorgado de acuerdo con la capacidad de pago.

17. Porcentaje de uso de cupo TC: Tasa de utilización de la línea rotatoria, o la cantidad de crédito que el prestatario está utilizando en relación con todo el crédito rotatorio disponible.

4.1.2 Procesamiento de las variables

Las variables independientes o variables de entrada son categorizadas según el tipo de información que arroja cada variable, estas se encuentran identificadas con una determinada nomenclatura misma que se muestra en la siguiente tabla:

Tabla 4. Identificación de las variables

Variable	Identificación	Tipo de la variable
Ingresos Anuales	annual_inc	Numérica
Ingresos Adicionales	total_rec_int	Numérica
Relación gastos-ingresos	Dti	Numérica
Monto de Crédito	loan_amnt	Numérica
Tasa de Interés	int_rate	Numérica
Calificación Buro de crédito	fico_range_high	Numérica
Última calificación Buro de crédito	last_fico_range_high	Numérica
Ultimo monto crédito	last_pymnt_amnt	Numérica
Morosidad 6 meses	inq_last_6mths	Numérica
Morosidad últimos 2 años	delinq_2yrs	Numérica
Total, cargos por mora	total_rec_late_fee	Numérica
Solicitudes negadas	pub_rec	Numérica
Número de créditos en la empresa	open_acc	Numérica
Hipotecas actuales	mort_acc	Numérica
Número total de créditos	total_acc	Numérica
Cupo rotativo TC	revol_bal	Numérica
Porcentaje de uso de cupo TC	revol_util	Numérica

4.1.3 Tratamiento de valores faltantes

En el procesamiento de los datos crediticios, se obtuvo variables que no disponían de información completa, que pueden ser consecuencia de errores en el registro, falla humana o

valores en blanco que no fueron otorgados por el cliente, en consecuencia, se debe depurar esta información para que no tenga errores los resultados del modelo crediticio. Según Kuhn y Johnson, los datos que faltan pueden ser imputados, en este caso, podemos usar la información de los predictores del conjunto de entrenamiento para, en esencia, estimar los valores de otros predictores (Max Kuhn y Kjell Johnson, 2013). La información con los datos faltantes se observa en la siguiente tabla:

Tabla 5. Datos faltantes

Variable	% Datos completos	% Datos faltantes
Ingresos Anuales	100%	0%
Ingresos Adicionales	100%	0%
Relación gastos-ingresos	95%	5%
Monto de Crédito	100%	0%
Tasa de Interés	99%	1%
Calificación Buro de crédito	93%	7%
Última calificación Buro d crédito	100%	0%
Ultimo monto crédito	99%	1%
Morosidad 6 meses	92%	8%
Morosidad últimos 2 años	100%	0%
Total, cargos por mora	100%	0%
Solicitudes negadas	85%	15%
Número de créditos en la empresa	93%	7%
Hipotecas actuales	77%	23%
Número total de créditos	82%	8%
Cupo rotativo TC	88%	12%
Porcentaje de uso de cupo TC	100%	0%

Para el tratamiento de los valores faltantes o perdidos se ha utilizado la imputación de datos empleando la técnica del vecino más cercano (KNN, K-Nearest Neighbor), que permite

imputar el valor faltante promediando los puntos cercanos (Yuli Sudriani, Foni Agus Setiawan y Abdul Hamid, 2020).

4.1.4 Análisis de correlación de las variables numéricas

Con el fin de encontrar problemas asociados a la multicolinealidad en las variables numéricas, es decir que exista una situación de correlación (Webster, 2016), se ejecuta una matriz de correlación entre las variables numéricas, determinando que existirá un problema de colinealidad si existe un índice de correlación superior a 0,8 (Gujarati, 2014).

Para el despliegue de esta matriz, se ha ocupado la librería ***pandas*** y ***seaborn***, específicamente el método corr() y heatmap(). El procedimiento usado para este proceso emplea los valores de correlación de ***Pearson***, que miden el grado de relación lineal entre cada par de variables. En la siguiente imagen se puede visualizar la codificación utilizada para este propósito.

```
cm_b = df.corr()
mask = np.triu(np.ones_like(cm_b, dtype=np.bool))
plt.figure(figsize=(10,10))
sns.heatmap(cm_b,mask=mask, xticklabels=df.columns,
            yticklabels=df.columns,annot=True,
            cmap = 'BrBG',fmt = ".1f")
```

Figura 23: Construcción de la matriz de correlación

Los valores resultantes de la matriz de correlación se obtienen al aplicar coeficiente de correlación de Pearson. Este índice permite medir el grado de covariación entre cada par de variables relacionadas linealmente. La fórmula general se expresa de la siguiente manera:

$$r_{xy} = \sum \frac{Z_x Z_y}{N}$$

Donde:

- x : es una de las variables a correlacionar.
- y : es la segunda variable
- z_x : Es la desviación estándar de la primera variable.
- z_y : Es la desviación estándar de la segunda variable.
- N : Es el número de observaciones.

En la **Figura 24** podemos revisar la matriz de correlación entre las diferentes variables que han sido propuesta para el modelo:

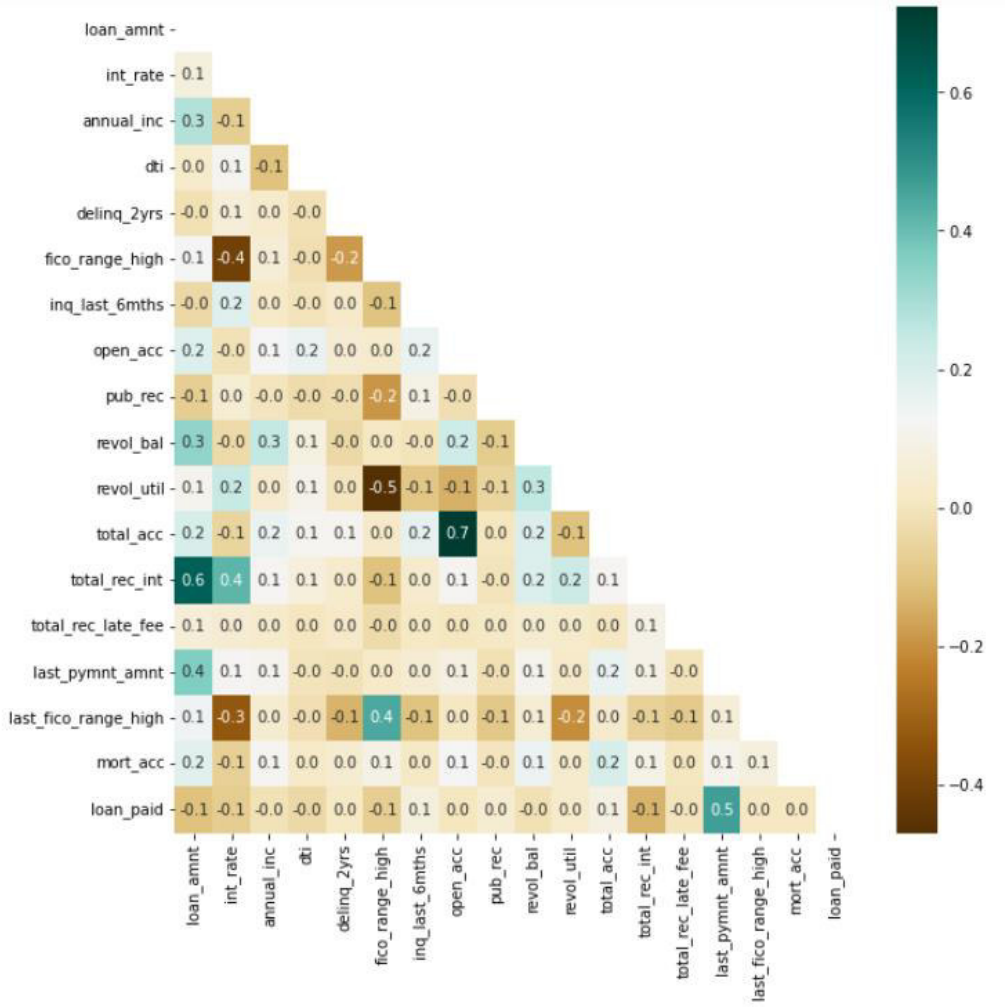


Figura 24: Matriz de correlación

No existe una correlación alta, que supere el 0,8 sugerido por Gujarati, por tanto, no es necesario reemplazar alguna de las variables numéricas. El valor más alto de correlación se encuentra entre el número de créditos en la empresa y el número total de créditos, ya que existe una correlación de 0.7, lo que se explica debido a medida que crecen los créditos en la empresa se incrementará el número de créditos totales del solicitante, esto es perfectamente entendible para el modelo de riesgo crediticio.

Con la información presentada del análisis de las variables numéricas se obtiene resultados favorables para el ingreso de información en el modelo de riesgo crediticio.

4.2 Modelado

En la literatura científica se encuentra un gran número algoritmos que pueden ser empleados para resolver el problema del pronóstico en la evaluación del riesgo crediticio (Siddharth Bhatore, Lalit Mohan y Raghu Reddy, 2020). Algunos de estos métodos son:

- Gradient Boosting (GB)
- Regresión Logística (LR)
- Árbol de Decisión (DT)
- Random Forest (RF)
- K-Nearest Neighbors (KNN)
- Artificial Neural Networks (ANN)
- Support Vector Machine (SVM)
- Multi Layer Perceptron (MLP).

Con base a estos modelos, se evalúa su uso en investigaciones con similares características con el fin de nutrir la investigación con los resultados obtenidos por otros investigadores. Para ello se emplea una búsqueda de artículos en el rango de fechas desde 2010 hasta 2021, con el propósito de contrastar los algoritmos de aprendizaje automático que han sido utilizados por diversos investigadores para abordar la evaluación del riesgo crediticio. A este tipo de investigación se le denomina documental (Nichols, 2010). Los resultados de esta investigación nos han arrojado los datos mostrados en la siguiente tabla:

Tabla 6. Comparación de modelos según la literatura

Referencia/Modelo	GB	LR	DT	RF	KNN	ANN	SVM	MLP	Total
(Aceituno, 2019)		X	X	X	X	X	X		6
(Grau, 2020)	X	X	X						3
(Ossa & Jaramillo, 2021)		X		X			X	X	4
(Iain Brown y Christophe Mues, 2012)		X		X	X	X	X		5
(Twala, 2010)		X	X		X	X			4
(Lean Yu, Xiao Yaoa, Shouyang Wanga y K. K. Lai, 2011)		X	X		X		X		4
(X. Zhang, Y. Yang y Z. Zhou, 2018)			X		X	X	X		4
(A.I. Marqués, V. García y J.S. Sánchez, 2012)		X	X		X	X	X		5
(Raquel Florez-Lopez y Juan Manuel Ramon-Jeronimo, 2015)		X	X		X	X	X		5
(Artem Bequé y Stefan Lessmann, 2017)		X	X		X	X	X		5
TOTAL	1	9	8	3	8	7	8	1	

Como se aprecia en la **Tabla 6**, el número mínimo de técnicas utilizadas en las investigaciones referenciadas se ubica en tres, y el máximo en 6. Si calculamos el número

medio de técnicas, en función del número de referencias, se puede decir que un número de 4 técnicas sería conveniente para la comparativa de estas.

Por otra parte, las técnicas que más han sido utilizadas en las investigaciones referenciadas nos indican que una buena elección seria:

a) Support Vector Machine
b) Regresión Logística
c) Árbol de Decisión
d) K-Nearest Neighbors

4.2.1 Métricas de evaluación de los modelos de Machine Learning

El uso de métricas para evaluar el comportamiento de un modelo predictivo es fundamental al momento de decidir cuál modelo es el más eficiente. En la literatura se presentas múltiples fórmulas que permiten determinar el rendimiento del modelo, tal es el caso del Accuracy, precisión, matriz de confusión, exhaustividad, entre otras ((Powers, 2020), (Amalia Luque, Alejandro Carrasco, Alejandro Martín y Ana de las Heras, 2019)). Según Sahli, el Acuracy es la métrica de evaluación más utilizada para la avaluación de modelos de clasificación en Machine Learning (Sahli, 2020). Considerando las métricas más utilizadas se propone emplear las siguientes (Hossin M. y Sulaiman M.N., 2015):

- Accuracy
- Precisión
- Recall
- F1 Score

A continuación, realizamos una breve explicación de cada una de ellas:

- Accuracy: también conocida como la métrica de exactitud, se encarga de medir el porcentaje de casos que el modelo de machine learning tuvo aciertos, es la métrica con mayor uso y en algunos casos llega a tener menor porcentaje de eficiencia (Martínez, 2020). La fórmula empleada por accuracy es la siguiente:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

  Donde:

  TP = Verdaderos Positivos

  TN = Verdaderos Negativos

  FP = Falsos Positivos

  FN = Falsos Negativos

- Precisión: es la métrica encargada de medir la calidad del modelo de machine learning (Martínez, 2020). Es definida como la tasa de observaciones positivas identificadas correctamente (Grau, 2020).

$$Precisión = \frac{TP}{TP + FP}$$

- Recall: conocida como exhaustividad y es la tasa de positivos reales que se identificó correctamente o la capacidad que tiene el modelo de identificar o predecir (Martínez, 2020). La fórmula es la siguiente:

$$Recall = \frac{TP}{TP + FN}$$

- F1 Score: es un indicador que evalúa la información obtenida en recall y precisión, toma en cuenta los falsos positivos y los falsos negativos (Ossa & Jaramillo, 2021). La fórmula es la siguiente:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4.2.2- Implementación de los algoritmos de machine learning

A continuación, se describirá los elementos más resaltantes que han sido considerados para el desarrollo de los algoritmos que han sido seleccionados en la sección anterior para la predicción del riesgo crediticio. Esta investigación se ha desarrollado empleando la herramienta Python3, en el entorno Anaconda, más concretamente con Jupyter Notebook. Se ha ocupado las librerías:

- Pandas
- Numpy
- Matplotlib
- Scikit-learn

La data que nos proporciona la empresa debe ser depurada para que pueda ser utilizada por algún algoritmo de aprendizaje automático. Para cumplir con este propósito se ha aplicado varios procesos como lo son:

- Revisión de la correlación de las variables proporcionadas por el modelo.

- Selección de las variables que más se adecuan con base al modelo de riesgo creditico utilizado.
- Imputación de valores faltantes o nulos: Dado que la data recolectada poseía datos faltantes se aplicó el método K-Nearest Neighbor para corregir la data.

4.2.3- División del Data set

Para realizar las pruebas de rendimiento del modelo es necesario dividir la data en:

- Porción de datos para el entrenamiento
- Porción de datos para las pruebas

Para realizar este proceso se ha empleado división aleatoria utilizando un 20% de los datos para prueba y 80% para el entrenamiento (Todd M. Martin, Paul Harten,Douglas M. Young, Eugene N. Muratov, Alexander Golbraikh, Hao Zhu y Alexander Tropsha, 2012).

En la siguiente tabla se puede revisar el número de observaciones para cada una de las porciones de datos resultantes.

Tabla 7: **División de los datos para entrenamiento y prueba**

Conjunto de datos	**Tamaño**
Set de entrenamiento	838860
Set de prueba	209715

4.2.4- Parámetros de los algoritmos

Para la implementación de los algoritmos se ha propuesto utilizar la menor cantidad de parámetros posibles de manera de obtener resultados aceptables. De acuerdo con la elección previa de los algoritmos, nos ocuparemos de estos atributos a continuación. Hay que aclarar

que la elección de estos parámetros no ha sido de forma arbitraria, se ha utilizado como criterio del ajuste estos hiperparámetros Grid Search (Rabiya Khalid y Nadeem Javaid, 2020). Cada uno de estos hiperparámetros se han seleccionado con base a los parámetros que recibe cada método de las librerías correspondientes que se encargan de implementar los algoritmos de machine learning propuestos. En la Tabla 8, se muestra los parámetros más relevantes que recibe cada uno de los métodos utilizados para su implementación:

Tabla 8: Hiperparámetros ajustados en cada modelo propuesto

Método	Hiperparámetros
SVC	Kernel, C y Gamma
LogisticRegression	Penalty y C
DecisionTreeClassifier	Criterion, max_depth, y ccp_alpha
KNeighborsClassifier	n_neighbors

Fuente: (scikit-learn.org, 2022)

Por otra parte, para afinar cada parámetro se ha empleado validación cruzada junto con la malla de valores asociados, en las siguientes figuras se puede ver el ajuste de estos parámetros.

Para el caso de los vectores de soporte, se ha fijado una malla con los parámetros que se van a ajustar con base a un rango de valores que pueden asumir cada uno de ellos de acuerdo al

algoritmo que se empleará. La **Figura 25**, se puede revisar los detalles de la malla y el código asociado.

```
# Grid de hiperparámetros
# ==============================================================================
param_grid = {'kernel':['linear','rbf'], 'C': [4.0,5.0,6.0,8.0], 'gamma': [1.0, 0.1, 0.01, 0.001]}

# Búsqueda por validación cruzada
# ==============================================================================
grid = GridSearchCV(
        estimator  = SVC(),
        param_grid = param_grid,
        scoring    = 'accuracy',
        n_jobs     = -1,
        cv         = 3,
        verbose    = 4,
        return_train_score = True
       )
```

Figura 25: Ajuste de parámetros para SVC

Los valores para estos hiperparámetros, luego de la ejecución de las pruebas, se pueden ver a continuación:

	param_C	param_gamma	param_kernel	mean_test_score	std_test_score	mean_train_score	std_train_score
8	5.0	1.0	linear	0.808120	0.009076	0.821811	0.007370
10	5.0	0.1	linear	0.808120	0.009076	0.821811	0.007370
12	5.0	0.01	linear	0.808120	0.009076	0.821811	0.007370
14	5.0	0.001	linear	0.808120	0.009076	0.821811	0.007370
30	8.0	0.001	linear	0.806917	0.007616	0.821216	0.007305

Figura 26: Mejores valores obtenidos para el SVC

Se aprecia que los mejores ajustes se alcanzan con un C=5.0, Gamma = 1.0 y un Kernel = linear.

Los ajustes para la regresión logística de los hiperparámetros han resultado en los siguientes valores: 'C': 0.05, 'penalty': 'l2'. El código usado se puede revisar en la **Figura 27Figura 27**.

```
# Grid de hiperparámetros
# ============================================================================
param_grid = {'penalty':['l1','l2'], 'C': [0.01,0.05,0.08,0.1,0.5]}

# Búsqueda por validación cruzada
# ============================================================================
grid = GridSearchCV(
        estimator  = LogisticRegression(),
        param_grid = param_grid,
        scoring    = 'accuracy',
        n_jobs     = -1,
        cv         = 3,
        verbose    = 1,
        return_train_score = True
       )
```

Figura 27: Ajuste de parámetros para Regresión Logística

Para el Árbol de decisión se emplea el código mostrado en la figura **Figura 28**, con la malla correspondiente.

```
# Grid de hiperparámetros
# ============================================================================
param_grid = {'criterion': ['gini', 'entropy'], 'max_depth':[5,10,15,20], 'ccp_alpha':[0.05,0.1,0.5,0.1]}

# Búsqueda por validación cruzada
# ============================================================================
grid = GridSearchCV(
        estimator  = DecisionTreeClassifier(),
        param_grid = param_grid,
        scoring    = 'accuracy',
        n_jobs     = -1,
        cv         = 3,
        verbose    = 1,
        return_train_score = True
       )
```

Figura 28: Ajuste de parámetros para Árbol de decisión.

Los valores resultantes para este ajuste resultaron en, 'ccp_alpha': 0.05, 'criterion': 'gini', 'max_depth': 5.

Finalmente, para KNeighborsClassifier, podemos revisar el código correspondiente en la **Figura 29**, y luego de la ejecución de la validación cruzada, los mejores valores obtenidos fueron: 'n_neighbors': 20.

```
# Grid de hiperparámetros
# ==============================================================================
param_grid = {'n_neighbors' : [15,18,20,25]}

# Búsqueda por validación cruzada
# ==============================================================================
grid = GridSearchCV(
        estimator  = KNeighborsClassifier(),
        param_grid = param_grid,
        scoring    = 'accuracy',
        n_jobs     = -1,
        cv         = 3,
        verbose    = 1,
        return_train_score = True
       )
```

Figura 29: Ajuste de parámetros para KNeighborsClassifier.

4.4.2.1- Support Vector Machine

Las máquinas de vectores de soporte se basan en el concepto de hiperplano y de los puntos de que definen el margen máximo de separación de las clases etiquetadas. Para el modelo propuesto se han empleado los siguientes atributos:

- Kernel: ***linear***, para hiperplanos lineales.
- C: Es el parámetro de regularización y se ha definido en 5.0.
- Gamma: Es el coeficiente de kernel, se ha ajustado en 1.0.

Los valores para estos parámetros han resultado de la validación cruzada por ajuste con malla de valores, explicado al principio de esta sección. El código empleado para ejecutar este modelo se puede revisar en la **Figura 30Figura 30**.

```
# ==============================================================================
# División de los datos en train y test
# ==============================================================================
X = df.drop(columns = 'loan_paid')
y = df['loan_paid']

X_train, X_test, y_train, y_test = train_test_split(
                                        X,
                                        y.values.reshape(-1,1),
                                        train_size   = 0.8,
                                        random_state = 1234,
                                        shuffle      = True
                                    )
# ==============================================================================
# Construcción del modelo
# ==============================================================================
model  = SVC(kernel='linear',C=5.0,gamma=1.0)

# ==============================================================================
# Entrenando
# ==============================================================================
model.fit(X = X_train, y = y_train)

# ==============================================================================
# Predicción
# ==============================================================================
predictions = model.predict(X_test)
```

Figura 30: Modelo de Vectores de soporte

Para el entrenamiento del modelo, lo primero que se debe hacer es dividir el set de datos en las porciones de prueba y entrenamiento, posteriormente se construye el modelo con base a los hiperparámetros ajustados previamente por la validación cruzada y la malla de valores. En el siguiente paso se entrena el modelo utilizando el set de entrenamiento, para luego aplicar las pruebas realizando la predicción con los datos de prueba, para posteriormente obtener las métricas del modelo. Estas métricas son las que nos permiten determinar qué tan eficiente es el modelo frente a los otros modelos propuestos.

4.2.4.2- Regresión Logística

La regresión logística es un modelo estadístico que nos permite estimar la relación existente entre un conjunto de variables cualitativas *Xi* y una variable cualitativa *Y*. En concreto, se refiere a un modelo lineal generalizado que utiliza una función logística como función de enlace. Parámetros utilizados en el modelo propuesto:

- Penalty: Indica el tipo de regularización a emplear, en este caso se ha utilizado 'L2' (regularización Ridge). La idea es minimizar el efecto de la correlación entre los atributos de entrada.
- C: Regularización del parámetro Penalty. Se definió en 0.05

En la **Figura 31**, se puede revisar el código utilizado para la construcción, entrenamiento y prueba del modelo de regresión logística. Al igual que la construcción del modelo de vectores de soporte, se ha empleado los hiperparámetros ajustados previamente por la validación cruzada y la malla de valores.

```
# ================================================================================
# División de los datos en train y test
# ================================================================================
X = df.drop(columns = 'loan_paid')
y = df['loan_paid']

X_train, X_test, y_train, y_test = train_test_split(
                                        X,
                                        y.values.reshape(-1,1),
                                        train_size   = 0.8,
                                        random_state = 1234,
                                        shuffle      = True
                                    )
# ================================================================================
# Construcción del modelo
# ================================================================================
model = LogisticRegression(C= 0.05, penalty='l2')

# ================================================================================
# Entrenando
# ================================================================================
model.fit(X = X_train, y = y_train)

# ================================================================================
# Predicción
# ================================================================================
predictions = model.predict(X_test)
```

Figura 31: Modelo de Regresión Logística

4.2.4.3- Árbol de Decisión

Los algoritmos basados en arboles de decisión corresponden a una estructura de datos jerárquica que se emplea para predecir la etiqueta asociada con una instancia $\boldsymbol{x}$ empezando en un nodo raíz de un árbol hasta una hoja.

- Criterion: Determina la calidad de una división (nuevas ramas). Para este caso se ha utilizado ‘gini’, prefiriendo la medida de impureza.
- max_depth: Maxima profunidad del árbol, definida en 5 niveles.
- ccp_alpha: Parámetro de complejidad utilizado para la poda de complejidad de costo mínimo. Definido en 0.05.

Para la construcción, entrenamiento y prueba se ha empleado el código que se muestra en la **Figura 32**. Los parámetros utilizados son los que resultaron del ajuste empleado validación cruzada y malla de valores.

```
# ================================================================================
# División de los datos en train y test
# ================================================================================
X = df.drop(columns = 'loan_paid')
y = df['loan_paid']

X_train, X_test, y_train, y_test = train_test_split(
                                        X,
                                        y.values.reshape(-1,1),
                                        train_size   = 0.8,
                                        random_state = 1234,
                                        shuffle      = True
                                    )
# ================================================================================
# Construcción del modelo
# ================================================================================
model = DecisionTreeClassifier(ccp_alpha= 0.05, criterion='gini', max_depth= 5)

# ================================================================================
# Entrenando
# ================================================================================
model.fit(X = X_train, y = y_train)

# ================================================================================
# Predicción
# ================================================================================
predictions = model.predict(X_test)
```

Figura 32: Modelo de Árbol de decisión

4.2.4.4- K-Nearest Neighbors

El algoritmo de vecino más cercano se fundamenta en la idea de memorizar el conjunto de entrenamiento y luego predecir la etiqueta de cualquier nueva instancia sobre la base de las

etiquetas de sus vecinos más cercanos en el entrenamiento. El parámetro más relevante corresponde a 'n_neighbors' y ha sido definido en 20.

Los valores para estos parámetros han resultado de la validación cruzada por ajuste con malla de valores, explicado al principio de esta sección. El código empleado para ejecutar este modelo se puede revisar en la **Figura 33**.

```
# ==============================================================================
# División de los datos en train y test
# ==============================================================================
X = df.drop(columns = 'loan_paid')
y = df['loan_paid']

X_train, X_test, y_train, y_test = train_test_split(
                                        X,
                                        y.values.reshape(-1,1),
                                        train_size   = 0.8,
                                        random_state = 1234,
                                        shuffle      = True
                                    )
# ==============================================================================
# Construcción del modelo
# ==============================================================================
model = KNeighborsClassifier(n_neighbors= 20)

# ==============================================================================
# Entrenando
# ==============================================================================
model.fit(X = X_train, y = y_train)

# ==============================================================================
# Predicción
# ==============================================================================
predictions = model.predict(X_test)
```

Figura 33: Modelo K-Nearest Neighbors

4.3- Evaluación

En esta sección se describe los elementos más relevantes del desarrollo, entrenamiento y prueba de los diferentes algoritmos de aprendizaje automático utilizados en esta investigación para abordar el problema de la asignación de créditos, considerando el modelo de credit scoring explicado anteriormente. Debemos recordar que estas propuestas ya incorporan el ajuste de los hiperparámetros indicados.

Por otra parte, dado que necesitamos evaluar el rendimiento de un problema de clasificación, podemos usar las métricas de precisión, recall, F1 y accuracy, que han sido explicada ampliamente en la sección 4.2.1.

4.3.1- Support Vector Machine

El algoritmo de soporte vectorial logra alcanzando un accuracy del 83%, lo que es un valor bastante bueno. Esto supone que se equivoca en un 27% de los casos. Hay que recordar que el ***acurracy*** mide el porcentaje de casos en el que el modelo de machine learning tuvo aciertos.

Tabla 9. Métricas del modelo de Support Vector Machine

Clase/Métrica	precisión	recall	f1-score	accuracy
No	0.73	0.92	0.82	83%
Si	0.93	0.76	0.84	

La métrica del área bajo la curva o AUC arroja un valor del 90%. El grafico de esta curva se observa en la Figura 34.

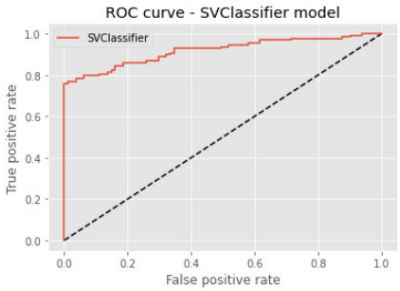


Figura 34: Curva ROC modelo Support Vector Machine.

Con la métrica de precisión es posible medir la calidad del modelo de en su tarea de clasificación, en otras palabras, la precisión es la respuesta a la pregunta ¿qué porcentaje de

los clientes que soliciten un crédito se les podrá otorgar o no el mismo? Para este modelo la precisión es del 83% en promedio.

La métrica de Recall nos permite conocer la cantidad de clientes que el modelo es capaz de identificar, es decir, es la respuesta a la pregunta ¿qué porcentaje de los clientes, a los cuales se les puede otorgar el crédito, somos capaces de identificar. En este caso se logra un porcentaje del 76%.

La métrica F1 asume que nos importa de igual forma la precisión y el Recall. En general intenta establecer un balance entre estas dos métricas. Se calcula una media armónica. Según esta métrica el modelo basado en vectores de soporte se logra un rendimiento del 84% para la predicción del otorgamiento del crédito, frente a un 82% para la no concesión.

4.3.2- Regresión Logística

El modelo alcanza un accuracy del 0.83% en la predicción de la clase correspondiente. Aunque para el 'No', no logra acertar en cerca del 28% del total de las observaciones de la prueba. En la siguiente tabla se ven las demás métricas.

Tabla 10. **Métricas del modelo de Regresión Logística**

Clase/Métrica	precisión	recall	f1-score	accuracy
No	0.72	0.96	0.82	83%
Si	0.97	0.73	0.83	

Respecto a la curva ROC, se registra un valor del AUC del 88%. El grafico de esta curva se observa en la Figura 35.

Figura 35: Curva ROC modelo Regresión Logística.

4.3.3- Árbol de decisión

El modelo basado en arboles de decisión muestra un rendimiento (accuracy) del 80% de exactitud. Siendo las métricas más relevantes las que se muestran a continuación.

Tabla 11. **Métricas del modelo de Árbol de decisión**.

Clase/Métrica	precisión	recall	f1-score	accuracy
No	0.68	0.99	0.80	80%
Si	0.99	0.67	0.80	

La curva ROC se puede revisar en la **Error! Reference source not found.**.

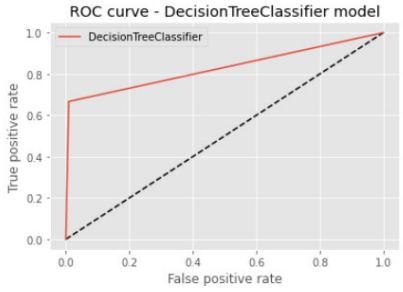


Figura 36: Curva ROC modelo Árbol de decisión.

Este modelo tiene una mayor dificultad de clasificar correctamente la clase del 'No', pues, se nota que no acierta en casi el 32% de los casos.

4.3.4- K-Nearest Neighbors

La métrica del accuracy nos muestra un 81% de certeza, con una mejora de 0.03, en la precisión de la predicción de la clase '**NO**' y a un 0.01 de cercanía con el modelo basado en regresión logística (véase la siguiente tabla).

Tabla 12. Métricas del modelo de K-Nearest Neighbors.

Clase/Métrica	precisión	recall	f1-score	accuracy
No	0.71	0.92	0.80	81%
Si	0.93	0.73	0.82	

En la Figura 37, se puede ver la curva ROC. Para este modelo la métrica AUC se ubica en 0.89.

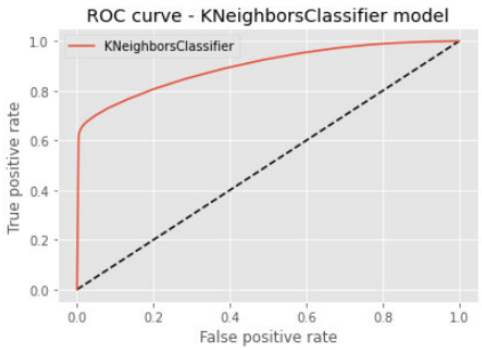


Figura 37: Curva ROC modelo K-Nearest Neighbors

4.4- Implantación

Una vez que se han realizado los ajustes del modelo la empresa decidirá cuándo pondrá el modelo estará disponibles en el entorno de producción. Una vez, aprobado su uso por la directiva, se dispondrá de las predicciones correspondientes que podrán ser utilizadas por otros departamentos de la empresa (Véase los anexos para detalles de capturas).

5. DISCUSIÓN DE LOS RESULTADOS

Dada la naturaleza de los datos y las características de las variables independientes, no sería justo hacer una comparativa con modelos de otros autores. Sin embargo, las referencias pueden ser un punto de partida para reforzar las hipótesis de algunos de estos autores. Con base a esto, se realizará una discusión de los resultados obtenidos en los modelos propuestos en esta investigación y se determinara, basados en las métricas empleadas, cual o cuales de estos pueden resultar más apropiados para resolver el problema planteado en los objetivos.

5.1- Comparativa de rendimiento con otros autores

En la siguiente tabla podemos revisar la métrica del ***accuracy*** alcanzadas por los modelos propuestos por algunos de los autores citados.

Tabla 13: Métricas de modelos propuestos por otros autores

Autor	**SVM**	**LR**	**DT**	**KNN**
(A.I. Marqués, V. García y J.S. Sánchez, 2012)	83%	83%	81%	80%
(Raquel Florez-Lopez y Juan Manuel Ramon-Jeronimo, 2015)	74%	74%	73%	79%
(Artem Bequé y Stefan Lessmann, 2017)	87%	86%	84%	83%
(Lean Yu, Xiao Yaoa, Shouyang Wanga y K. K. Lai, 2011)	83	82%	**72%**	**77%**
(Aceituno, 2019)	92%	96%	94%	82%
(Grau, 2020)	--	86%	85%	--
(Ossa & Jaramillo, 2021)	**58%**	**59%**	--	--
(Oscar Pucha, 2022)	***83%***	***83%***	***80%***	***81%***

Considerando los limites inferiores reportados por los autores citados, podemos decir que los modelos propuestos, en esta investigación, superan este umbral, lo que nos hace suponer que los resultados obtenidos están justificados por los trabajos relacionados. Sin embargo, como ya lo hemos indicado, hay factores que pueden afectar esta comparativa (i.e, el número de predictores utilizados en cada modelo, el balance de las clases, etc.).

5.1- Comparativa de los modelos propuestos

En general, los modelos propuestos logran un porcentaje de precisión de más del 80%, lo podríamos considerar como bastante buenos, para la predicción del riesgo creditico. Sin embargo, hay que mencionar que, a pesar de mostrar un buen rendimiento, se les complica la predicción de las clases minoritarias. Esto se puede explicar debido al desbalance que muestra la data utilizada para el entrenamiento y prueba en el desarrollo de estos.

En la Figura 38, se puede ver cómo ha sido el comportamiento en la precisión de cada uno de los modelos para cada clase (Si - otorgar el crédito, No – No otorgar).

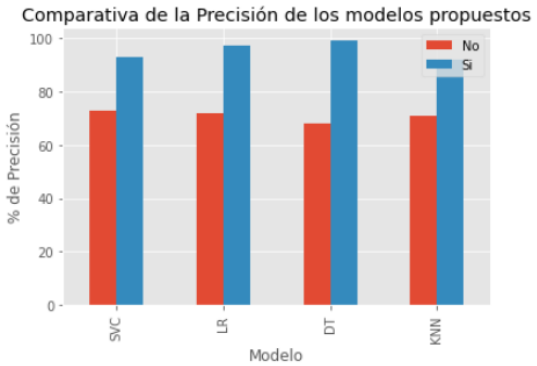


Figura 38: Métrica de precisión para cada modelo (clases SI,NO)

Se aprecia como la relación de precisión entre las clases es inversamente proporcional, es decir, con el aumento del acierto en la clase ***SI*** disminuye la precisión en la clase ***NO***. Se puede suponer que los modelos tienen un comportamiento muy similar en este aspecto.

Por otro lado, la curva ROC nos dice qué tan bueno puede distinguir el modelo entre las dos clases a utilizar en la clasificación de los clientes para otorgar o no el crédito.

El comparativo en las métricas encontradas en la aplicación de los modelos de machine learning se muestran en la tabla 13.

Tabla 14. **Comparación métricas modelo machine learning**

	Accuracy	**Precision**	**Recall**	**F1 Score**
Support Vector Machine (SVM)	83%	83%	84%	83%
Regresión Logística (LR)	83%	84%	85%	83%
Árbol de Decisión (DT)	80%	83%	83%	80%
K-Nearest Neighbors (KNN)	81%	82%	82%	81%

En el indicador de exactitud (accuracy) el mejor puntaje lo obtiene el modelo de support vector machine y el modelo de regresión logística, lo que indica que son mejores modelos para predecir los casos de aciertos. Como ya se ha señalado antes, el accuracy mide el porcentaje de casos en los que el modelo ha acertado (tanto para el otorgamiento del crédito como para su rechazo). Considerando esta métrica los resultados favorecen a los modelos que se han mencionado.

En el indicador de precisión el mejor resultado lo obtiene el modelo de regresión logística, estableciendo que estos modelos tendrían mejor calidad en su sistema de predicción en los

casos de aciertos y no aciertos. Recordamos que esta métrica nos permite evaluar el modelo con base a la repercusión que tiene los falsos positivos sobre el pronóstico.

En el indicador de recall, el mejor resultado lo obtiene el modelo de árbol de decisión y el modelo de regresión logística, que indica que son mejores modelos para medir los aciertos en la concesión de créditos. Es una medida de la media de los casos positivos que fueron pronosticados correctamente por el clasificador, sobre todos los casos positivos de los datos.

El indicador F1 Score el mejor resultado son support vector machine y regresión logística, lo que muestra una consecuencia de los indicadores de precisión y recall. En este caso, la media armónica nos da una mejor idea del rendimiento del modelo.

A nivel global, los cuatro modelos empleados tienen resultados similares y son óptimos para predecir el riesgo de crédito, sin embargo, muestra menores resultados el modelo árbol de decisión y K-Nearest Neighbors frente a modelos de regresión logística y support vector machine.

CONCLUSIONES

En este trabajo, se logró plantear el modelo de predicción de riesgo de crédito basado en mecanismos de machine learning. Los modelos utilizados fueron Regresión Logística, Árbol de Decisión, K-Nearest Neighbors y Support Vector Machine; en todos los modelos se encontraron resultados satisfactorios y pueden usarse en la predicción de riesgo. La diferencia radica en su efectividad, el modelo de Regresión logística y Support Vector Machine demostraron ser más eficaces en la predicción que sus contrapartes propuestas. Esta conclusión coincide con otras investigaciones relacionadas, donde los resultados presentados son similares a este documento.

El objetivo planteado para la fundamentación teórica se cumplió ya que se diseñó un esquema teórico y conceptual que partió con una reseña de la inteligencia artificial y machine learning. En este apartado se detalló conceptualmente la concepción de los algoritmos en los modelos de predicción de riesgo lineales y árbol de decisión. Adicionalmente en el apartado de marco teórico se abordó información referente al riesgo de crédito, lo cual permitió conocer la fundamentación del proceso crediticio en las instituciones financieras. De esta manera se concluye que la investigación tiene respaldo teórico y conceptual para entender las razones y circunstancias que llevaron a aplicar los modelos de predicción mencionados.

El objetivo de analizar la información para alimentar los modelos de predicción se cumple en su totalidad con el análisis de las variables numéricas y categóricas, en el caso de las variables numéricas se realiza un completo análisis estadístico con datos de la desviación estándar, máximo y mínimos, para verificar la calidad de los datos y poder alimentar los modelos con información relevante. En el caso de información faltante se utilizó la media y

la moda para completar los datos en la muestra de estudio. La principal conclusión referente a este aspecto determina que los datos utilizados cumplen con las condiciones estadísticas para alimentar los modelos.

Para desarrollar los modelos de predicción se utilizó una metodología de desarrollo ágil con el fin de planificar, ordenar y ejecutar la información en base a un esquema sistematizado de actividades. En el caso de la evaluación de los modelos de predicción se utilizó cuatro indicadores, estos fueron: accuracy, precisión, recall y F1 Score; que son indicadores empleados con el fin de medir la efectividad y bondad de los resultados que presentan los modelos de predicción.

BIBLIOGRAFÍA

A.I. Marqués, V. García y J.S. Sánchez. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, 10244-10250.

A.I. Marqués, V. García y J.S. Sánchez. (2012). Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications*, 10916-10922.

Aceituno, M. (2019). *Modelo predictivo de análisis de riesgo crediticio usando machine learning del sector microfinanciero.* Lima: Universidad Comillas.

Alfred DeMaris. (1992). *Logit Modeling: Practical Applications.* Sage Publications, inc.

Amalia Luque, Alejandro Carrasco, Alejandro Martín y Ana de las Heras. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 216-231.

Artem Bequé y Stefan Lessmann. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications*, 42-53.

B. Venkatesh y J. Anuradha . (2019). A Review of Feature Selection and Its Methods . *CYBERNETICS AND INFORMATION TECHNOLOGIES*, 3 - 26.

Bank for International Settlements. (octubre de 2013). *Marco Regulador Internacional para Bancos - Basilea III.* Obtenido de Marco regulador internacional para bancos - Basilea III: http://www.bis.org/bcbs/basel3_es.htm

Baviera, T. (2016). Técnicas para el análisis del sentimiento en Twitter: Aprendizaje Automático Supervisado. *Dígitos*, 33 - 50.

Bernal, C. (2018). *Metodología de la Investigación.* Bogota: McGraw Hill.

Bidaurrazaga, A. (2019). *Estudio de diferentes modelos de redes neuronales para el desarrollo de clasificador de tareas.* Bilbao: Universidad del Pais Vasco.

Borrero, D., & Bedoya, O. (2020). Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia articial. *UIS Ingenierías*, 37 - 52.

Byanjankar, A. (2017). *Predicting Credit Risk in Peer-to-Peer Lending with Survival Analysis.* Honolulu: Serie de simposios IEEE 2017 sobre inteligencia computacional (SSCI).

Cela, G., & Cuenca, J. (2018). *Propuesta de modelo de machine learning para la evaluación de riesgo de crédito utilizando algoritmos de predicción para la Cooperativa de Ahorro y Crédito La Merced.* Cuenca: Universidad del Azuay.

Chih-Fong Tsai, Yu-Feng Hsu y David C. Yen. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, 977-984.

Chun-Ling Chuang y Rong-Ho Lin. (2009). Constructing a reassigning credit scoring model. *Expert Systems with Applications*, 1685-1694.

Cisneros, D. (2016). *Gestión integral de riesgos.* Perú: Universidad ESAN.

Delgado, M. (2016). *La Inteligencia Artificial.* Granda: Universidad de Granada.

Duarte, M. (2018). Los Algoritmos en la vida cotidiana: desafíos estratégicos. *Revista Monitor Económico de Baja California*, 1915 - 1925.

E. Altman, R. Haldeman y P. Narayanan. (1977). Zeta Analysis. *Journal of Banking and Finance*, 29-54.

Galicia, M. (2013). Los enfoques de riesgo de crédito. *Instituto del Riesgo Financiero*, 20 - 82.

Gallardo Arancibia, J. (2014). *Metodología para el Desarrollo de Proyectos en Minería de Datos.* Buenos Aires: Universidad de Palermo.

Gómez, M. (2014). *Introducción a la metodología de la investigación científica.* Buenos Aires: Brujas.

Grau, J. (2020). *Machine Learning y Riesgo de Crédito.* Madrid: Pontificia Universidad de Comillas.

Gujarati, D. (2014). *Econometría.* Mexico DF: McGraw Hill.

Hill, R. (2016). What algorithm is. *Philosophy and Technology*, 35 - 39.

Hongmei Chena y Yaoxin Xiang. (2017). The Study of Credit Scoring Model Based on Group Lasso. *Procedia Computer Science*, 677-684.

Hossin M. y Sulaiman M.N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 01–11.

Iain Brown y Christophe Mues. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 3446-3453.

Jones, H. (2019). *El aprendizaje automático y su implicación en inteligencia artificial, minería de datos y Big Data.* Madrid: McGraw Hill.

Kotsiantis, S. (2017). Supervised machine learning. A review of classification techniques. *Informatica*, 249 - 268.

Labanda, X. (2017). *Implementación de un modelo de evaluación de riesgo de crédito y mercado, inherentes al financiamiento de proyectos inmobiliarios con recursos de la Mutualista Pichincha.* Quito: UCE. Recuperado el 13 de diciembre de 2020, de http://www.dspace.uce.edu.ec/bitstream/25000/15503/1/T-UCE-0005-CEC-002.pdf

Laínez Fuentes, J. (2014). *Desarrollo de Software AGIL.* México DF: McGraw Hill.

Lean Yu, Xiao Yaoa, Shouyang Wanga y K. K. Lai. (2011). Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection. *Expert Systems with Applications*, 15392-15399.

Martínez, J. (09 de octubre de 2020). *www.iartificial.net*. Obtenido de www.iartificial.net: https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/

Max Kuhn y Kjell Johnson. (2013). *Applied Predictive Modeling.* New York: Springer.

McCarthy, R. (2020). *El Método Agile.* México DF: Mc Graw Hill.

Moine, J. (2014). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. *Revista Argentina de Ciencias de la Computación*, 931- 938.

Molina, K. (2014). *La Cultura en la Era Digital.* San Jose: Universidad de Costa Rica.

Monasterio Astobiza, A. (2017). Ética algorítmica: Implicaciones éticas de una sociedad cada vez más gobernada por algoritmos. *Ética de datos, sociedad y ciudadanía*, 185 - 217.

Navarro, A., Fernández, J., & Morales, J. (2013). Revisión de metodologías ágiles para el desarrollo de software. *Prospectiva*, 30 - 39.

Nichols, B. (2010). *Introduction to Documentary.* Bloomington: Indiana University Press.

Nisha Arora y Pankaj Deep Kaur. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing*.

Nooritawati Md Tahir, Aini Hussain, Salina Abdul Samad, Khairul Anuar Ishak y Rosmawati Abdul Halim. (2006). Feature Selection for Classification Using Decision Tree. *2006 4th Student Conference on Research and Development.* Malaysia: IEEE.

Ossa, W., & Jaramillo, V. (2021). *Machine Learning para la estimación de riesgo de crédito en cartera de consumo.* Medellin: EAFIT.

Pelaez, I. (2016). Modelos de regresión lineal simple y regresión logística. *Revista Seden*, 195 - 214.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*.

Rabiya Khalid y Nadeem Javaid. (2020). A survey on hyperparameters optimization algorithms of forecasting models in smart grid. *Sustainable Cities and Society*.

Raquel Florez-Lopez y Juan Manuel Ramon-Jeronimo. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. *Expert Systems with Applications*, 5737-5753.

Rivillas, C., & Reina, W. (2016). Estimación del riesgo de crédito en empresas del sector real en Colombia. *Estudios Gerenciales*, 169 - 190.

Rodríguez M., Piñeiro Carlos y De Llano P. (2014). Determinación del riesgo de fracaso financiero mediante la utilización de modelos paramétricos de inteligencia artificial, y de información de auditoría. *Estudios de Economía*.

Rodriguez, J., & Miñano, M. (2017). *Desarrollo de una aplicación informática basada en un modelo de Machine Learning para mejorar la evaluación de préstamos crediticios.* Trujillo: Universidad Privada del Norte.

Rojo Aceituno, M. (2019). *Modelo predictivo de análisis de riesgo crediticio usando machine learning en el sector microfinanciero.* Puno: Universidad Nacional del Altiplano.

Romillo, M. (2019). *Modelo predictivo de análisis de riesgo crediticio usando machine learning en una entidad del sector microfinanciero.* Lima: Universidad del Altiplano.

Ross, J., Westerfield, A., & Jaffe, G. (2014). *Finanzas Corporativas* (Tercera ed.). Bogotá: McGraw Hill.

Rouhiainen, L. (2018). *Inteligencia Artificial.* Barcelona: Planeta.

Sahli, H. (2020). An Introduction to Machine Learning. En D. Laffly, *TORUS 1 – Toward an Open Resource Using Services* (págs. 61-74). ISTE Ltd and John Wiley & Sons, Inc. .

Sánchez, G. (2018). *Gestión del riesgo de crédito en las entidades financieras dentro del marco normativo Basilea IV.* Madrid: Pontificia Universidad de Comillas.

Sandoval, L. (2018). Algoritmos de aprendizaje automático para análisis y predicción de datos. *Revista Tecnológica* , 36 - 40.

scikit-learn.org. (01 de 02 de 2022). *scikit-learn.org*. Obtenido de https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

Score, F. (09 de 01 de 2022). *FICO Score*. Obtenido de https://www.myfico.com/credit-education/credit-scores/

Siddharth Bhatore, Lalit Mohan y Raghu Reddy. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology* , 111–138.

Superintendencia de Bancos y Seguros. (2004). *Normas Generales para las Instituciones del Sistema Financiero.* Quito: Resolución Nº JB-2004-631 de 22 de enero de 2004.

Taoufikallah, A. (2018). *La metodología Action Research.* Sevilla: Escuela Técnica Superior de Ingenieros de Sevilla.

Tejero, E. (2020). Algoritmos. El totalitarismo determinista que se avecina. *Revista de pensamiento estratégico y seguridad* , 85 - 102.

Todd M. Martin, Paul Harten,Douglas M. Young, Eugene N. Muratov, Alexander Golbraikh, Hao Zhu y Alexander Tropsha. (2012). Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *Journal of Chemical Information and Modeling*, 2570 - 2578.

Turing, A. (2008). *Inteligencia Artificial: La tecnología del futuro.* San Jose: Universidad de Costa Rica.

Twala, B. (2010). Multiple classifier application to credit risk assessment. *Expert Systems with Applications*, 3326-3336.

Villareal, D. (2013). *Adopción de una metodología ágil para proyectos de software.* México DF: Universidad Europea Miguel de Cervantes.

Webster, A. (2016). *Estadistica aplicada a los negocios.* Mexico DF: McGraw Hill.

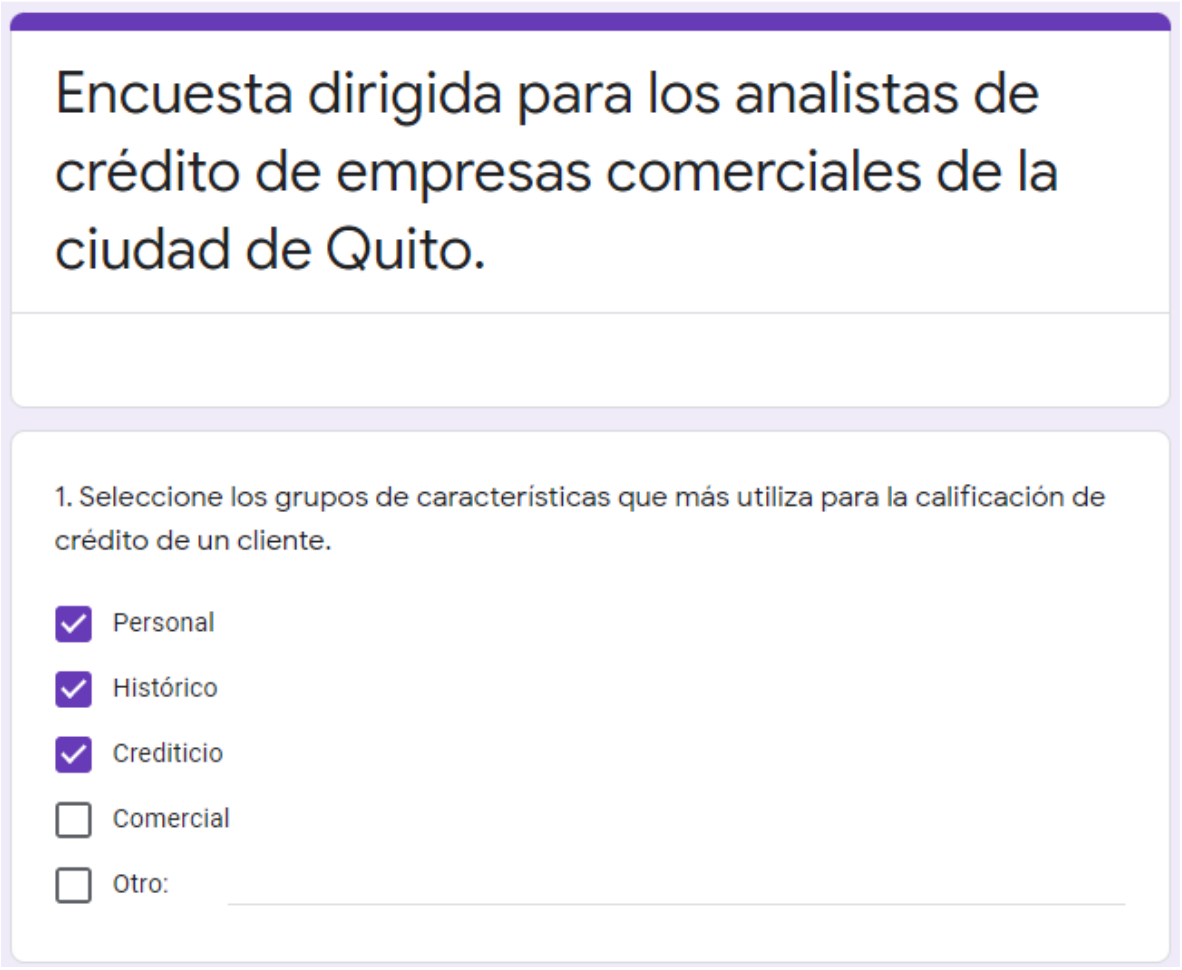
X. Zhang, Y. Yang y Z. Zhou. (2018). A novel credit scoring model based on optimized random forest. *8th Annual Computing and Communication Workshop and Conference* (págs. 60-65). IEEE.

Yuli Sudriani, Foni Agus Setiawan y Abdul Hamid. (2020). Comparison of kNN and Iterative Imputation Approach for Missing Data Value of Online Water Quality Monitoring System in Lake Maninjau. *(Jurnal Online Informatika*.

Zhou, H. (2017). A brief of introductory to weakly supervised learning. *National Sciencia Reviem*, 44 - 53.

ANEXOS

Anexo 1. Encuesta



2. Entre las características personales del cliente, a cuál le da más peso en la decisión de conceder un crédito.

- [x] Ingresos Mensuales
- [ ] Egresos Mensuales
- [x] Ingresos Adicionales
- [ ] Relación gastos-ingresos
- [x] Nivel educativo
- [x] Domicilio
- [x] Estado civil
- [x] Dependientes
- [ ] Otro: _____

3. Entre las características históricas del cliente, a cuál le da más peso en la decisión de conceder un crédito.

- [x] Calificación Buro de crédito
- [x] Morosidad
- [ ] Créditos aprobados, negados
- [x] Hipotecas
- [x] Créditos obtenidos
- [ ] Cupo TC
- [x] Tiempo de trabajo
- [ ] Tiempo de cliente
- [ ] Otro: _____

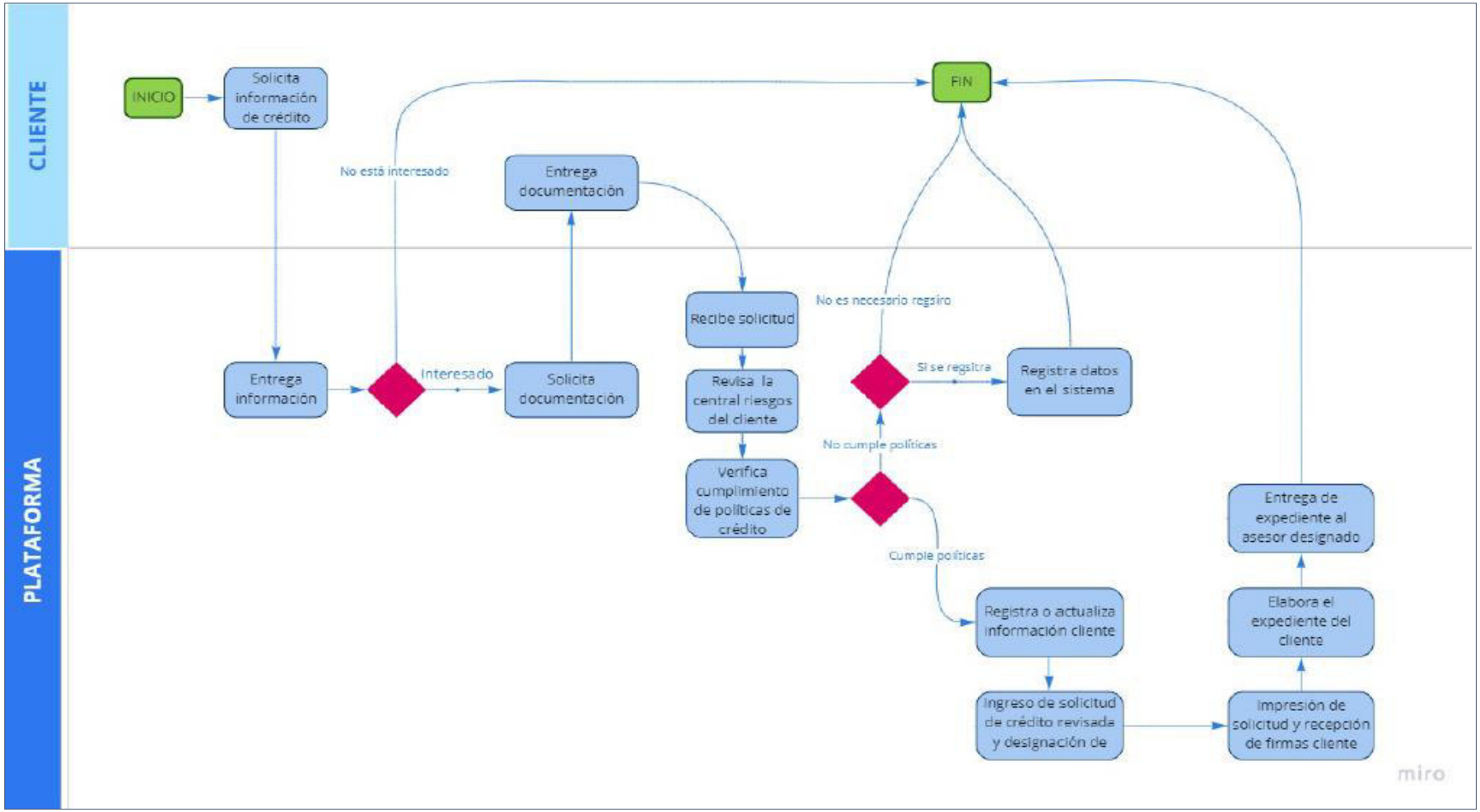
4. Entre las características históricas del cliente, a cuál le da más peso en la decisión de conceder un crédito.

- [x] Monto de Crédito
- [x] Tasa de Interés
- [x] Tipo de crédito
- [ ] Tiempo del crédito
- [ ] Saldos en la cuenta
- [ ] Tipo de cuenta
- [x] Temporalidad
- [ ] Información crediticia
- [ ] Otro: _____

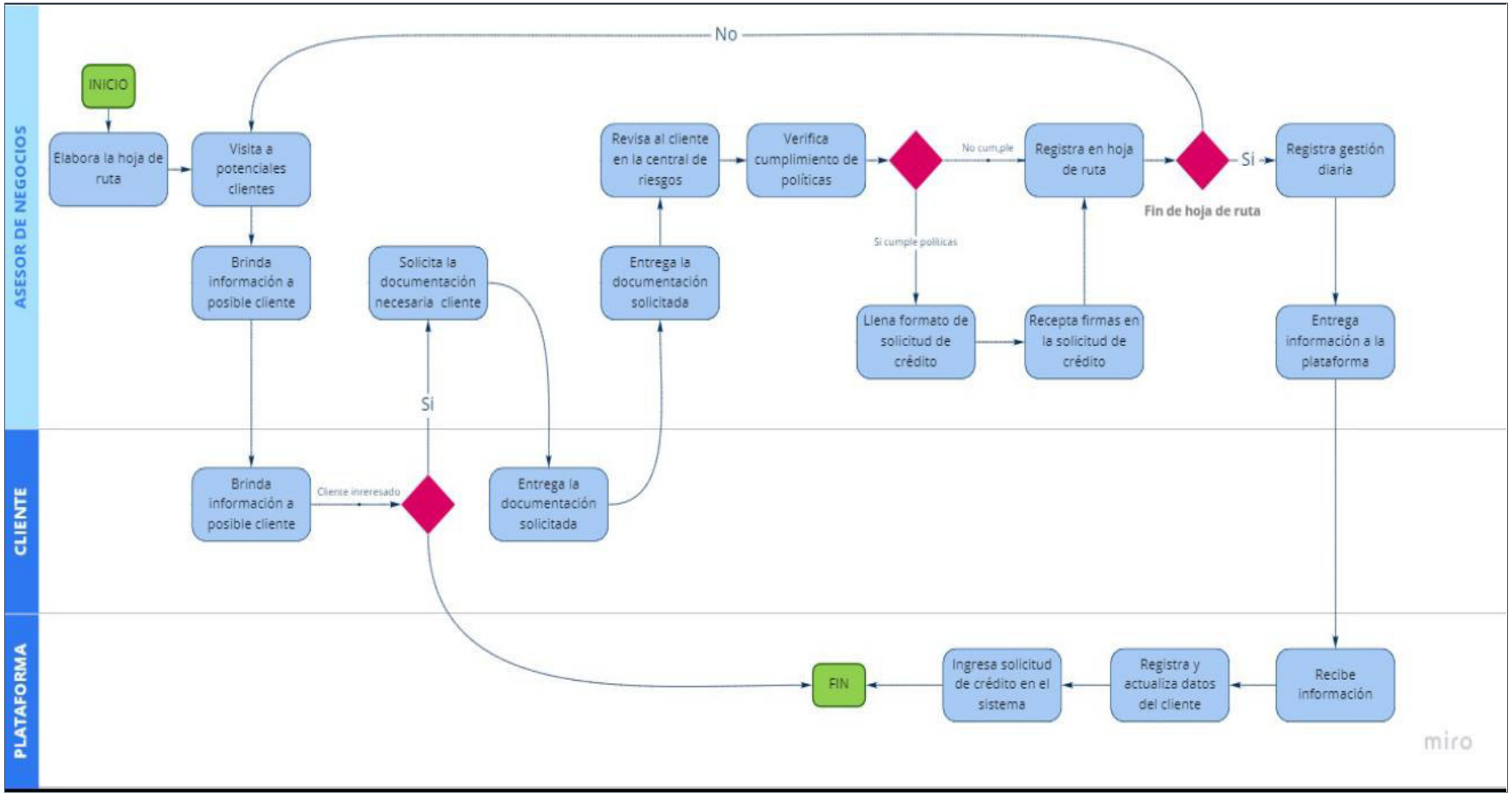
5. Que herramientas utiliza para la evaluación de crédito.

- [x] Análisis basado en la experiencia
- [x] Aplicativos ofimáticos
- [ ] Programas especializados
- [ ] Inteligencia artificial
- [ ] Otro: _____

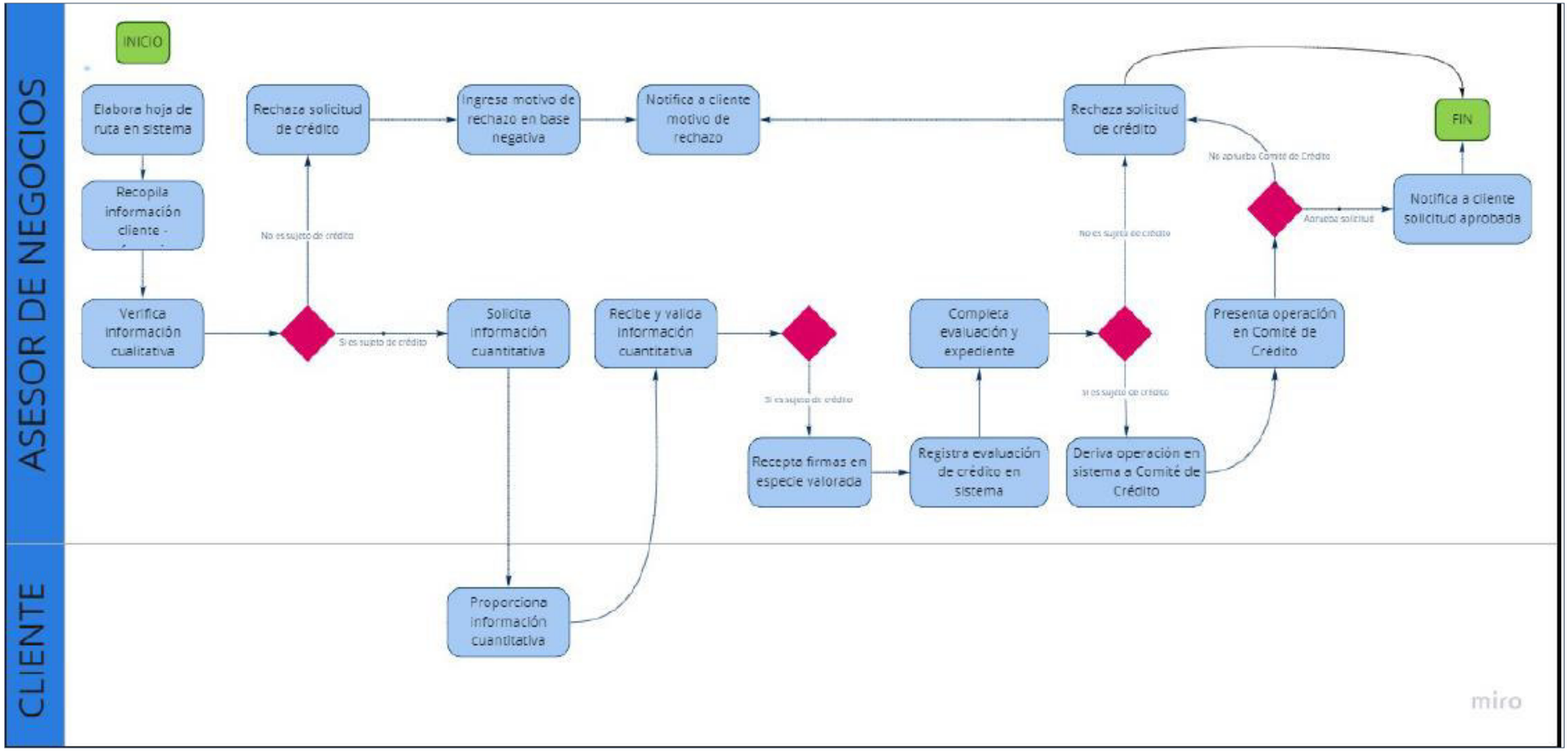
Anexo 2. Proceso de captación pasiva de créditos



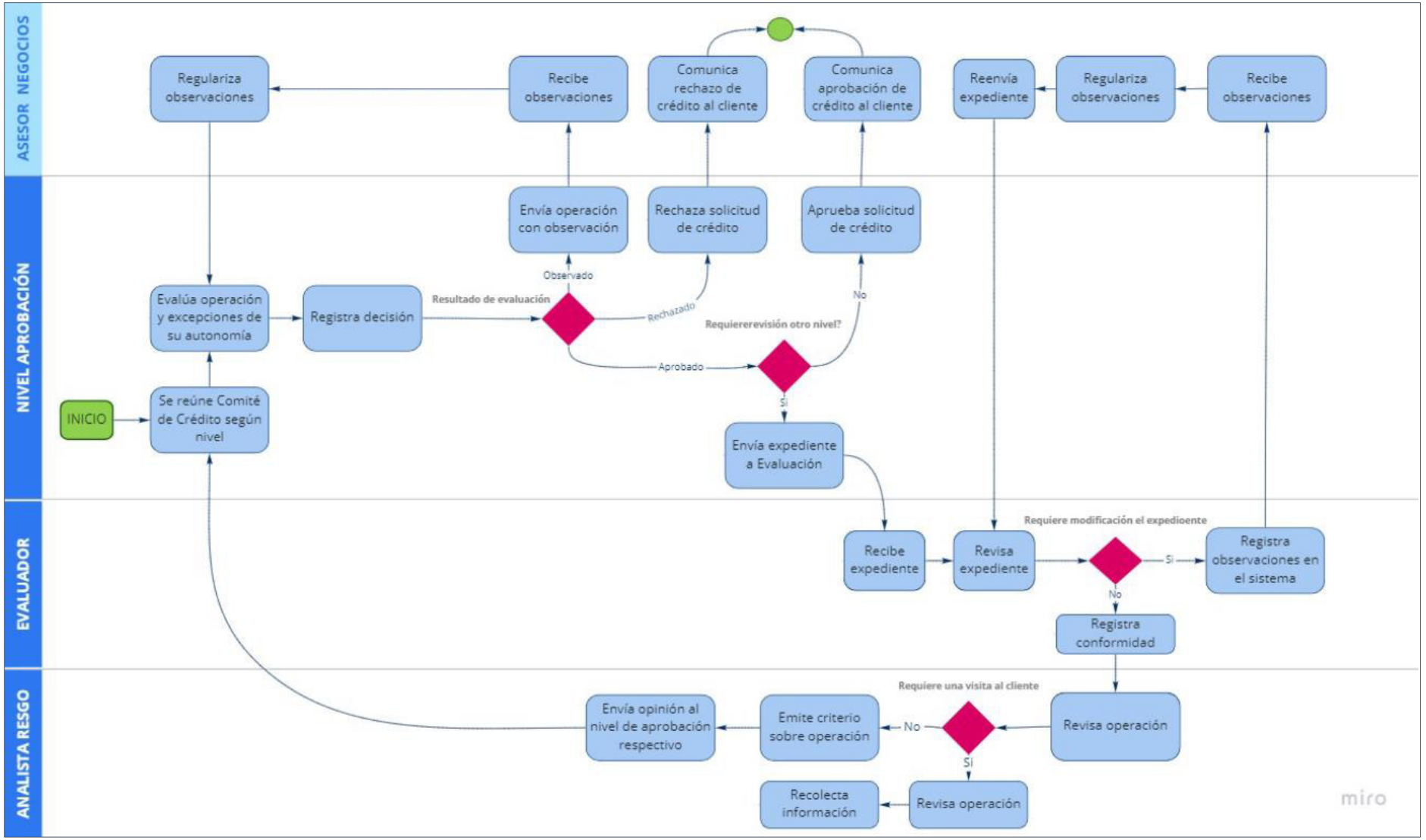
Anexo 3. Proceso de captación activa de créditos



Anexo 4. Proceso de evaluación de créditos



Anexo 5. Proceso de aprobación de créditos



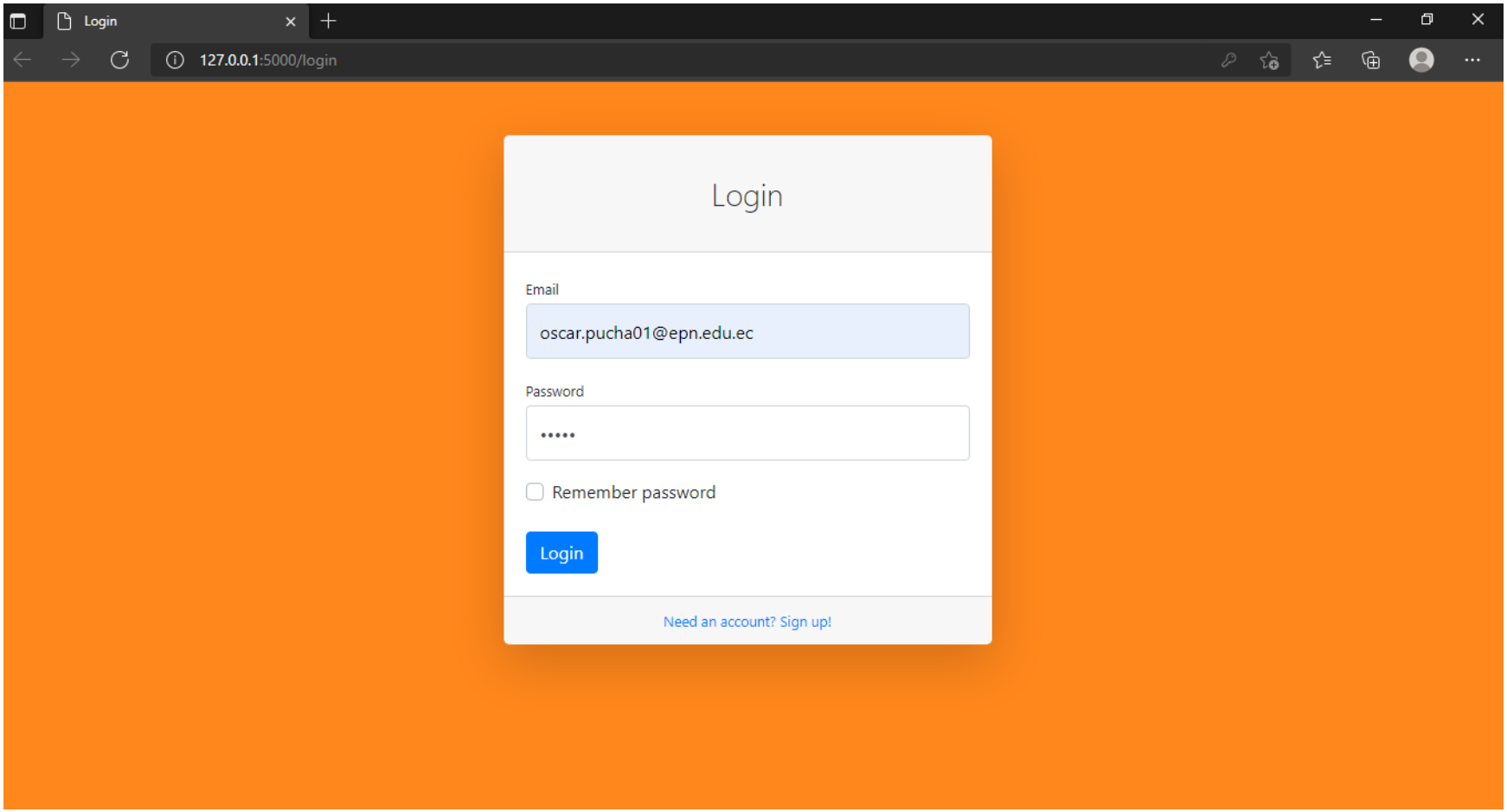
Anexo 6. Calculo de estadisticas

	loan_amnt	int_rate	annual_inc	dti	delinq_2yrs	fico_range_high	inq_last_6mths	open_acc	pub_rec	revol_bal
count	545077.000000	545077.000000	545077.000000	545077.000000	545077.000000	545077.000000	545077.000000	545077.000000	545077.000000	545077.000000
mean	14939.353935	12.203037	69033.977824	18.706086	0.330062	702.284514	0.510299	11.923137	0.207727	15410.430433
std	8296.753480	4.545690	17747.463040	6.087481	0.910289	32.751545	0.813831	5.402804	0.571208	14631.158234
min	1000.000000	5.310000	40903.000000	7.740000	0.000000	664.000000	0.000000	1.000000	0.000000	0.000000
25%	8800.000000	8.460000	54000.000000	13.740000	0.000000	679.000000	0.000000	8.000000	0.000000	6783.000000
50%	14000.000000	11.550000	67000.000000	18.400000	0.000000	694.000000	0.000000	11.000000	0.000000	12053.000000
75%	20000.000000	14.650000	82375.000000	23.480000	0.000000	719.000000	1.000000	15.000000	0.000000	19818.000000
max	40000.000000	30.990000	107996.000000	31.050000	58.000000	850.000000	6.000000	93.000000	52.000000	605063.000000

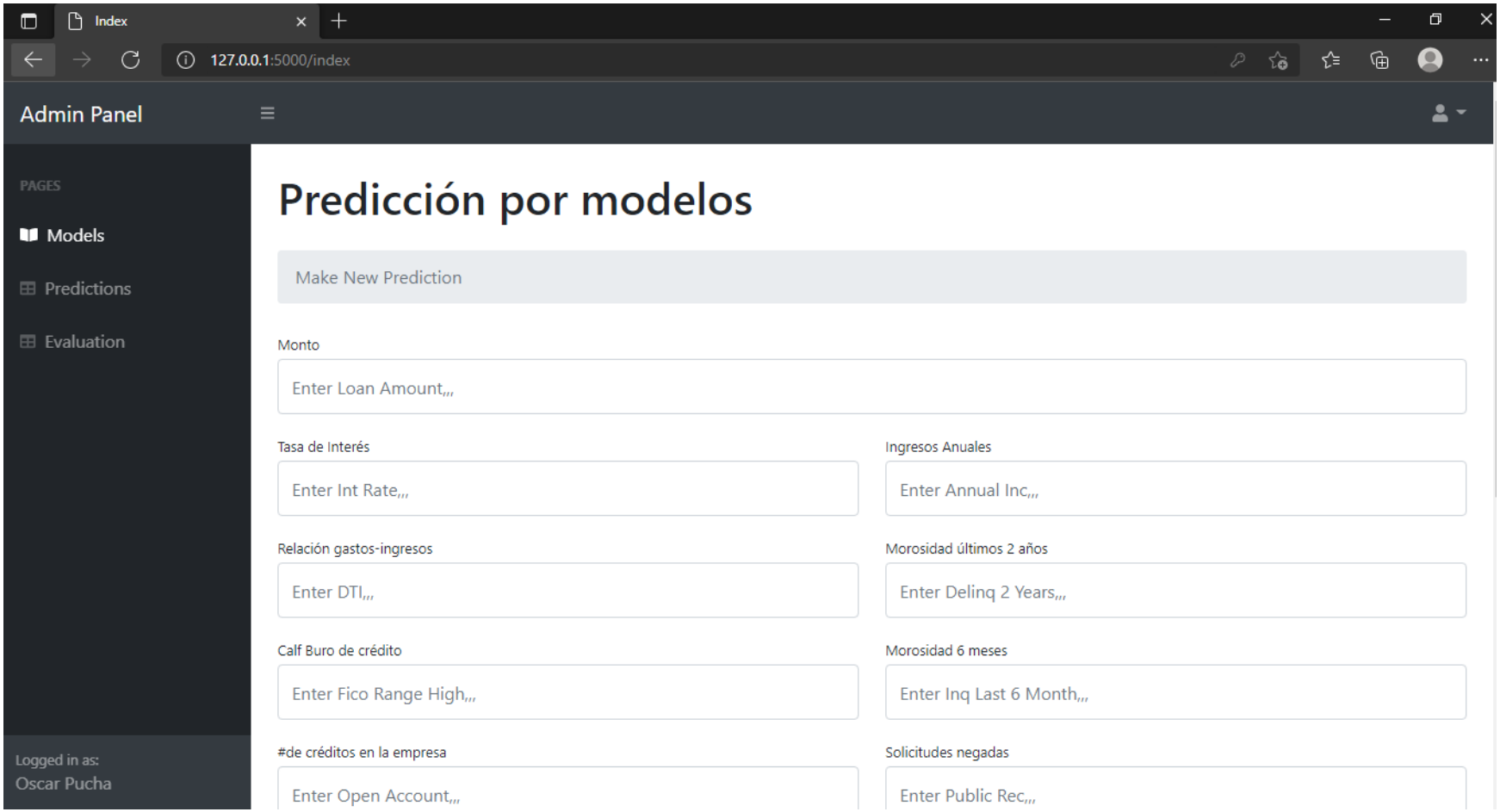
pub_rec	revol_bal	revol_util	total_acc	total_rec_int	total_rec_late_fee	last_pymnt_amnt	last_fico_range_high	MORTGAGE	loan_paid
545077.000000	545077.000000	545077.000000	545077.000000	545077.000000	545077.000000	545077.000000	545077.000000	545077.000000	545077.000000
0.207727	15410.430433	50.620652	24.855116	2486.033612	0.846051	4223.698333	705.630082	0.516619	0.597222
0.571208	14631.158234	24.016579	11.413491	2613.585179	8.762628	6310.337685	57.424634	0.499724	0.490457
0.000000	0.000000	0.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	6783.000000	32.400000	17.000000	793.740000	0.000000	339.790000	674.000000	0.000000	0.000000
0.000000	12053.000000	50.300000	23.000000	1599.290000	0.000000	721.220000	709.000000	1.000000	1.000000
0.000000	19818.000000	68.900000	31.000000	3212.240000	0.000000	6268.810000	744.000000	1.000000	1.000000
52.000000	605063.000000	191.000000	156.000000	27309.350000	1098.360001	42192.050000	850.000000	1.000000	1.000000

Anexo 7. Front End de la Aplicación Web para el analisis de riesgo crediticio

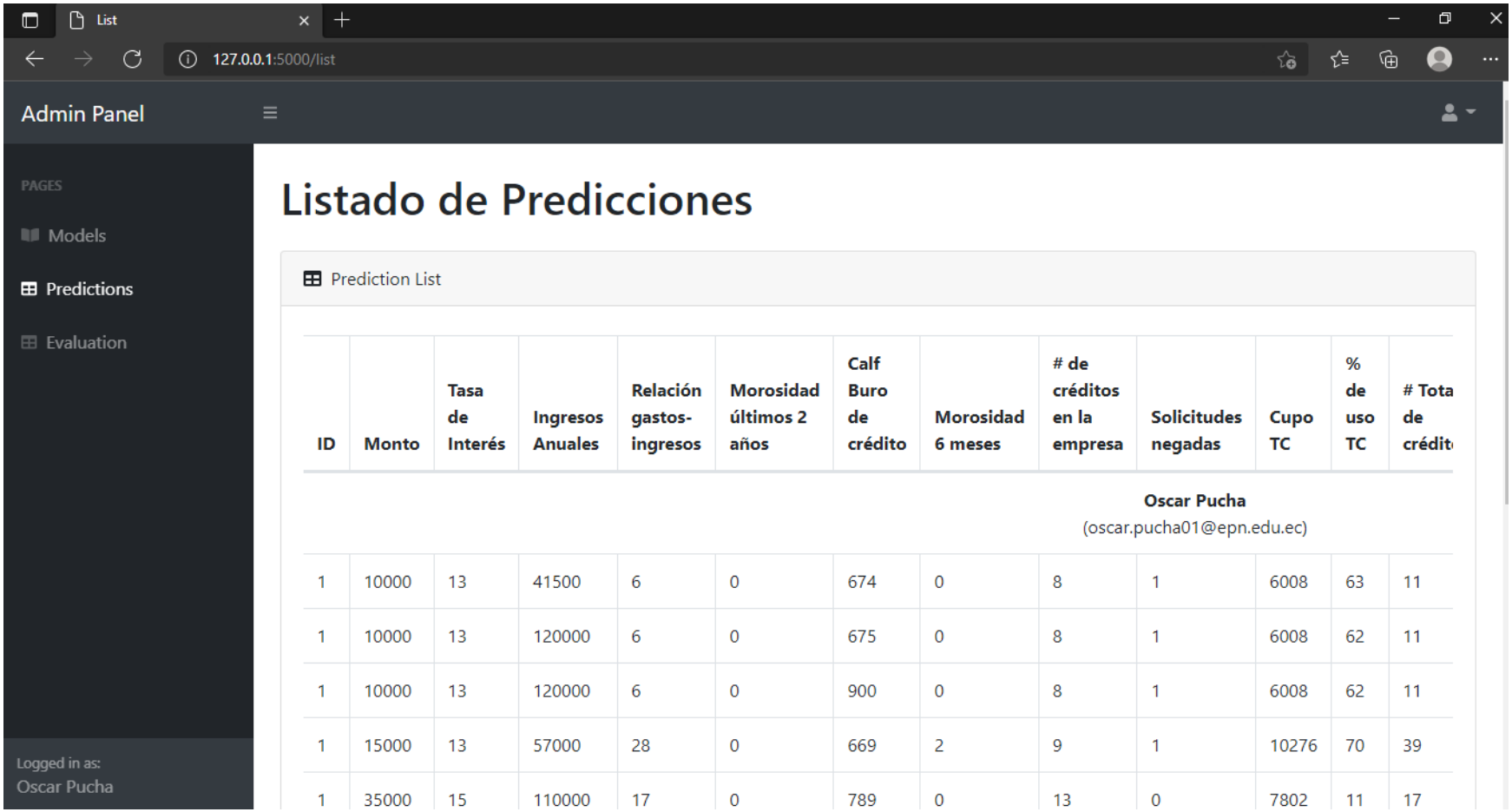
Pagina de Ingreso



Anexo 8. Módulo de prediccion de riesgo crediticio

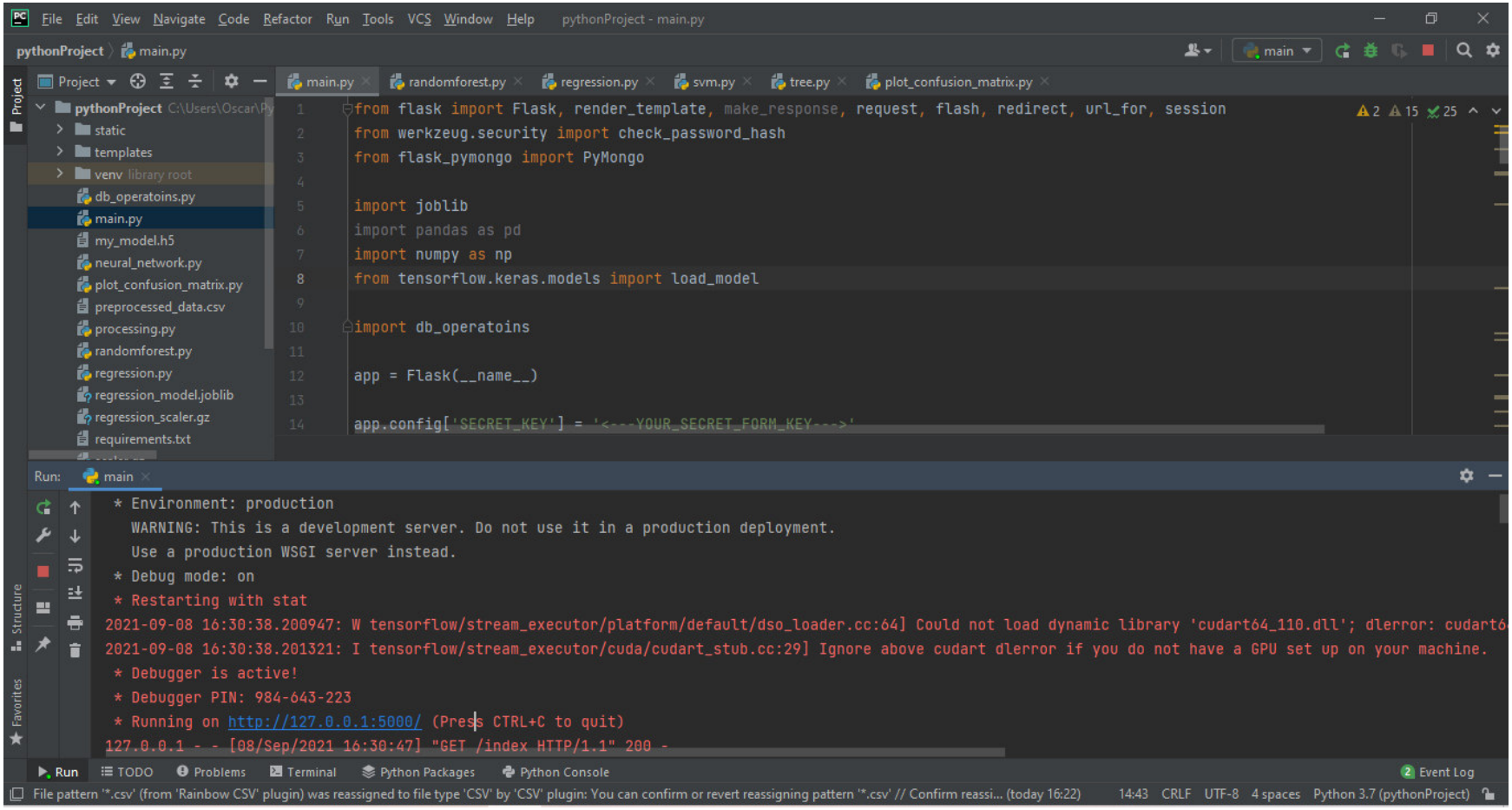


Anexo 9. Listado de predicciones

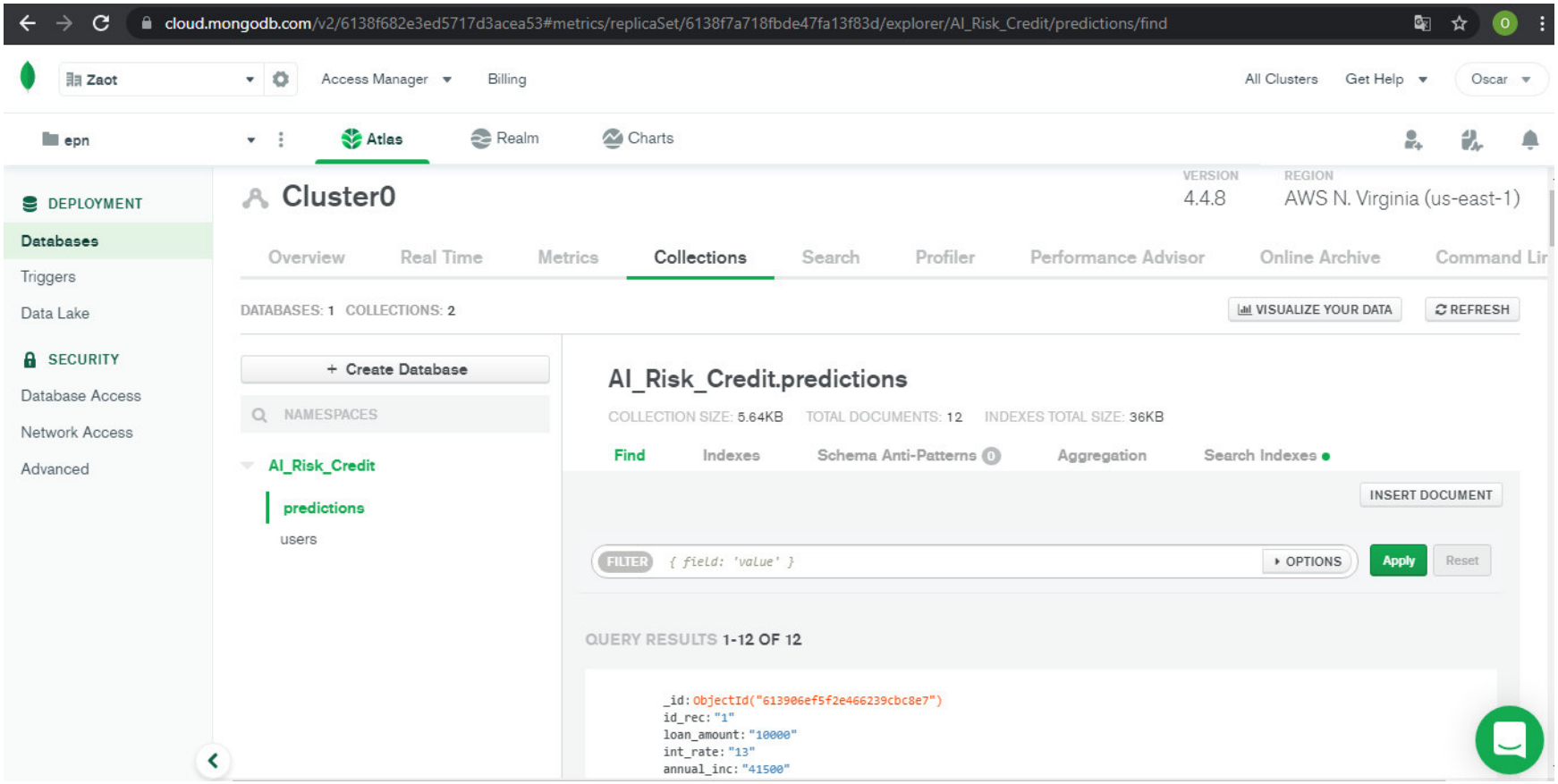


Anexo 10. Back End de la Aplicación Web para el analisis de riesgo crediticio

PyCharm Community Edition 2021.1.3 x64



Anexo 11. MongoDB Atlas



Anexo 12. Modelo Python: Regresión Logística (RL)

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = MinMaxScaler()

X_train= scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

print(X_train.shape)
print(X_test.shape)

# # building the model
#
# joblib.dump(scaler, 'regression_scaler.gz')
scaler = joblib.load('regression_scaler.gz')

# model = LogisticRegression()
# model.fit(X_train, np.ravel(y_train))

# joblib.dump(model, "regression_model.joblib")
model = joblib.load("regression_model.joblib")

predictions = model.predict(X_test)
```

Anexo 13. Modelo Python: Random Forest (RF)

```
df = df_accepted.sample(frac=0.8,random_state=42)
print(len(df))

df = df_accepted.copy()

X = df.loc[:, df.columns != 'loan_paid'].values
y = df.loan_paid.values

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = MinMaxScaler()

X_train= scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

print(X_train.shape)
print(X_test.shape)

# # building the model

joblib.dump(scaler, 'random_forest_scaler.gz')
scaler = joblib.load('random_forest_scaler.gz')

model = RandomForestClassifier()
model.fit(X_train, np.ravel(y_train))
```

Anexo 14. Modelo Python: Support Vector Machine (SVM)

```
X = df.loc[:, df.columns != 'loan_paid'].values
y = df.loan_paid.values
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = MinMaxScaler()

X_train= scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

print(X_train.shape)
print(X_test.shape)

# # building the model
#
joblib.dump(scaler, 'SVC_scaler.gz')
scaler = joblib.load('SVC_scaler.gz')

model = SVC()
model.fit(X_train, np.ravel(y_train))
#
joblib.dump(model, "SVC_model.joblib")
model = joblib.load("SVC_model.joblib")

predictions = model.predict(X_test)
```

Anexo 15. Modelo Python: Árbol de decisión (AD)

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = MinMaxScaler()

X_train= scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

print(X_train.shape)
print(X_test.shape)

# # building the model
joblib.dump(scaler, 'tree_scaler.gz')
scaler = joblib.load('tree_scaler.gz')

model = DecisionTreeClassifier()
model.fit(X_train, np.ravel(y_train))
#

# loading saved model
joblib.dump(model, "tree_model.joblib")
model = joblib.load("tree_model.joblib")


# predicting the Test data
predictions = model.predict(X_test)
```

Anexo 16. Java Script: Pagina web de evaluación de Riesgo crediticio