UNIVERSIDAD PRIVADA DEL NORTE

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA DE SISTEMAS COMPUTACIONALES

“DESARROLLO DE UNA APLICACIÓN INFORMÁTICA BASADA EN UN MODELO DE MACHINE LEARNING PARA MEJORAR LA EVALUACIÓN DE PRÉSTAMOS CREDITICIOS”

Tesis para optar el título profesional de:

Ingeniero de Sistemas Computacionales

Autor(es):
Rodríguez Castillo, Jorge Junior
Miñano Ochoa, Milagros Madeleine

Asesor:
Mg. Juan Orlando Salazar Campos

Trujillo – Perú
2017

APROBACIÓN DE LA TESIS

El asesor y los miembros del jurado evaluador asignados, **APRUEBAN** la tesis desarrollada por los Bachilleres Jorge Junior Rodríguez Castillo y Milagros Madeleine Miñano Ochoa, denominada:

"DESARROLLO DE UNA APLICACIÓN INFORMÁTICA BASADA EN UN MODELO DE MACHINE LEARNING PARA MEJORAR LA EVALUACIÓN DE PRÉSTAMOS CREDITICIOS"

Mg. Juan Orlando Salazar Campos
ASESOR

Ing. Paúl Alexander Quiñones Martínez
JURADO
PRESIDENTE

Ing. José Alberto Gómez Ávila
JURADO

Ing. José Humberto Vásquez Pereyra
JURADO

DEDICATORIA

Dedicamos este trabajo de investigación a nuestros padres, hermanos (as) y familiares quienes nos apoyaron en cada momento de nuestras vidas, ya que gracias a ellos culminamos nuestra carrera profesional y el desarrollo de esta investigación.

AGRADECIMIENTO

A nuestros maestros de la Universidad Privada del Norte por brindarnos sus enseñanzas durante nuestra etapa de estudiantes, también a nuestro querido asesor de tesis Mg. Juan Orlando Salazar Campos por el tiempo y la dedicación brindada para lograr culminar de manera exitosa el presente trabajo de investigación.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El presente trabajo de investigación está enfocado en el estudio de un modelo de machine learning para desarrollar una aplicación informática, que permita mejorar la evaluación de préstamos crediticios brindando un mejor análisis de la rentabilidad y el riesgo crediticio; la cual sea usada por la empresa financiera que por motivos de privacidad de sus datos llamaremos Financiera Nuestro Crédito.

El problema radica en determinar como una aplicación informática basada en un modelo de machine learning contribuye a mejorar la evaluación de préstamos crediticios.

Para solucionar dicha problemática se desarrolló una aplicación informática basada en un modelo de regresión logística que permite realizar la evaluación y predicción de préstamos crediticios por medio de una interfaz sencilla, donde se ingresan características principales como; el monto solicitado, la tasa de interés, los plazos del crédito, el estado civil y la edad del solicitante. El modelo de regresión logístico está dividido en cuatro algoritmos principales; un algoritmo para el escalamiento de datos, un algoritmo para predicción denominado función sigmoidal, una función para reducir el costo del modelo y el algoritmo de optimización de la gradiente de descenso. Esta aplicación informática bajo el uso del modelo de regresión logística logrará aumentar el porcentaje de dinero ganado, disminuir la cantidad de dinero perdido y disminuir el tiempo promedio para la aprobación de préstamos crediticios.

Los resultados del trabajo de investigación indican que con el desarrollo de esta aplicación informática basada en el modelo de regresión logística se logra aumentar el porcentaje de dinero ganado, se logra disminuir la cantidad de dinero perdido con la evaluación de los préstamos crediticios y el tiempo promedio empleado para aprobar un préstamo crediticio. Además, la eficacia de esta aplicación informática brinda un porcentaje aceptable de predicción para la empresa financiera. Por eso se concluye, que dicha aplicación informática es de gran utilidad para la Financiera Nuestro Crédito.

ABSTRACT

The present research work is focused on the study of a machine learning model to develop an informatic application, which allow to improve the evaluation of credit loans by providing a better analysis of the profitability and credit risk; which is used by the financial company that for reason of privacy of its data we will call Financiera Nuestro Crédito.

The problema lies in determining how an informatic application based on a machine learning model contributes to improving the evaluation of credit loans.

To solve this problema was developed an informatic application based on a logistic regression model which allows to perform the evaluation and prediction of credit loans through a simple interface, where the user put some principal features such as; el amount requested, interest rate, credit terms, marital status and the age of the applicant. The logistic regression model is divided in four main algorithms, the first one is a function for scaling data, the second one is an algorithm for prediction that is called sigmoid function, the third one is a function to reduce the cost of the model and the last one is the algorithm of optimization called the gradient descent. This informatic application under the use of the logistic regression model will achieve to increase the percentage of earned money, decrease the amount of lost money and decrease of the time average for the approval of credit loans.

The results of the reasearch work indicate that with the development of this informatic application based on a logistic regression model, the percentage of earned money is increased, the amount of lost money is reduced by the evaluation of the credit loans and the average time spent to approve a credit loan is reduced, too. Besides that, the effectiveness of this informatic application give an aceptable percentage of prediction for the financial company. For these reason, it is concluded that this informatic application is very useful to Financiera Nuestro Crédito.

CAPÍTULO 1. INTRODUCCIÓN

1.1. Realidad problemática

En la actualidad es evidente el crecimiento de los datos y de la automatización de muchas actividades que se realizan en el día a día. Por eso, la tecnología se ha convertido en un aspecto muy importante para las organizaciones y sus participantes, contribuyendo en la toma de decisiones (Gonzales Andrés, 2014).

Las entidades financieras han sido las pioneras en utilizar Data Mining y Machine Learning, lo han aplicado básicamente en el núcleo de su negocio, la financiación. Muchas veces cuando un cliente desea solicitar un préstamo, se le solicita determinada información como edad, estado civil, nivel de ingresos, domicilio, etc. la cual queda almacenada sin ser explotada en toda su magnitud. (Gómez Pedro, 2016).

Las entidades financieras al otorgar créditos o préstamos influyen de cierta manera en la economía, por eso, cuando no hay acceso al crédito, el consumo de las familias y la inversión de las empresas debe financiarse con los ingresos de cada período, generando inconvenientes cuando los ingresos son muy variables (BCRP, 2009). En el informe del Banco Interamericano de Desarrollo (2004), menciona que otra de las razones de la importancia de la actividad crediticia es que, al no existir un sistema de intermediación financiera eficiente para ejecutar la distribución de los recursos, se limita el emprendimiento de proyectos rentables restringiendo el crecimiento económico. Según Añez Manfredo (2001): "Es necesario conocer a través de un análisis cuidadoso los estados financieros del cliente, análisis de los diversos puntos tanto cualitativos como cuantitativos que en conjunto permitirá tener una mejor visión sobre el cliente y la capacidad para poder pagar dicho crédito."

En el ámbito internacional, las entidades financieras han combinado la tecnología con su principal actividad: evaluación crediticia. Existen varias startups con soluciones que permiten disminuir, gracias a algoritmos de scoring como máquina de soporte de vectores (SVM), redes neuronales, redes bayesianas y K-mean, los niveles de morosidad de bancos y empresas de concesión de préstamos o prestar dinero con menos riesgo que con otros modelos de evaluación (BBVA, 2015). Según una encuesta virtual realizada por la empresa Belatrix, 40% de organizaciones dicen estar investigando activamente sobre temas de Machine Learning, mientras 12% ya había comenzado una iniciativa de aprendizaje automático (Alex Robbio, 2016). MoneyMan, un sitio web de préstamos online, indica que la compañía Lending Club utiliza machine learning para construir modelos de créditos efectivos y Kueski, una startup con sede en Guadalajara, se dedica a dar microcréditos por Internet teniendo como base machine learning y procesos de información que ayudan a decidir si dar o no los créditos (María Fernanda Díaz, 2017).

En el Perú, aún no existen empresas financieras que estén utilizando machine learning para mejorar el proceso de evaluación de préstamos crediticios. En nuestro país el número de personas que presentan incumplimientos en sus pagos de deudas financieras es muy alto, así lo dio a conocer el gerente de proyectos de Financiera Edyficar, Gaby Cárdenas: "Hay entre 80 mil a 100 mil clientes castigados por año en todos los segmentos (créditos a pymes, de consumo, etc.) del sistema financiero. Son clientes que no pagan, a pesar de hacer todos los esfuerzos por cobrarles y ubicarlos, no cumplen con la obligación, la SBS exige que cubras esas pérdidas" (Comercio, 2014).

En Trujillo, algunos bancos y cajas utilizan modelos scoring para la evaluación de préstamos. Según el Ingeniero José Vásquez P., la Caja Trujillo es una de las entidades financieras que utiliza un modelo estadístico de scoring.

En la entrevista realizada a un analista de crédito (ANEXO 01) indicó que a ellos se les asigna una cartera de clientes, para que los visiten y les informen de los créditos disponibles; en el caso que los clientes acepten, los analistas inician la evaluación verificando que el solicitante no reporte atrasos de deudas en la central de riesgos de la superintendencia de Banca y Seguros. Seguidamente, confirman los datos de los documentos solicitados; los cuáles son, documento de identidad (DNI), recibo de luz, agua, o telefonía fija, declaración anual de impuestos a la renta, etc. Otra de las opciones que tienen los clientes es acercarse directamente a la entidad financiera para solicitar su préstamo, ahí deberán seguir y cumplir con ciertas reglas. Estos procedimientos pueden durar aproximadamente un día o dos.

De nuestra realidad problemática podemos deducir que la evaluación crediticia tiene deficiencias en nuestro país, pues mantener un procedimiento de evaluación crediticia como el descrito anteriormente ocasiona que el tiempo se prolongue para realizar una evaluación crediticia y el riesgo de incumplimiento de éste sea mayor.

1.2. Formulación del problema

¿De qué manera una aplicación informática basada en un modelo de machine learning mejora la evaluación de préstamos crediticios?

1.3. Justificación

Según lo planteado en el presente proyecto, la evaluación de préstamos crediticios hoy en día es un punto vital en el crecimiento económico tanto para entidades financieras como para el país. Las entidades financieras por medio del uso de una aplicación basada en un modelo de machine learning se verán beneficiadas, ya que esta aplicación informática analiza de manera directa los datos necesarios para brindar un mejor resultado, permitiendo ahorrar tiempo en la evaluación de los préstamos crediticios por el uso de algoritmos de machine learning y ahorrar costos en contratación de personal y uso de recursos, ya que la aplicación

informática realiza el proceso de evaluación con una mayor precisión para reducir índices de morosidad.

El presente proyecto de investigación servirá como base para futuros temas de investigación en lo que respecta al tema de machine learning, ya que cuenta con información importante sobre algoritmos y técnicas para el uso de machine learning, no solo en temas de evaluación de préstamos crediticios sino también para cualquier tema que requiera el uso de este campo de estudio.

1.4. Limitaciones

- No contar con una base de datos completa para realizar el modelo basado en machine learning; por lo cual, se pretende fijar un periodo de tiempo con la información que se cuenta para desarrollar el modelo en base a esta información.

1.5. Objetivos

1.5.1. Objetivo general

Determinar la mejora de la evaluación de préstamos crediticios mediante una aplicación informática basada en un modelo de machine learning.

1.5.2. Objetivos específicos

- Aumentar el porcentaje de dinero ganado por los préstamos crediticios clasificados.
- Disminuir la cantidad de dinero perdido por préstamos crediticios clasificados.
- Disminuir el tiempo promedio en días empleado para aprobar un préstamo crediticio clasificado.
- Aumentar el porcentaje de préstamos crediticios clasificados de manera correcta por medio del uso de la aplicación informática con el modelo de machine learning.

CAPÍTULO 2. MARCO TEÓRICO

2.1. Antecedentes

(Aboobyda & Tarig, 2016). Desarrollo de modelo predictivo para riesgo de crédito en los bancos con uso de minería de datos. Khartoum, Sudán. En esta investigación se demostró que el uso de minería de datos y machine learning ayudan a reducir el riesgo de asignar un crédito bancario; mediante la técnica de minería de datos realizan la extracción de la información necesaria para definir el modelo y luego realizar la fase de aprendizaje con machine learning. Se utilizaron los algoritmos como j48, redes bayesianas y bayes ingenuo en la plataforma Weka, así se logró corregir y minimizar los errores en la fase de aprendizaje obteniendo un modelo limpio y con una precisión más acertada sobre el riesgo de crédito. Además, Jafar y Mohammed hacen énfasis en la importancia de que los datos extraídos sean normalizados para tener un modelo predictivo válido. Los resultados obtenidos en esta investigación en la categoría de exactitud de clasificación dicen que con el uso del algoritmo j48 se logró un 78.38%, redes bayesianas 77.47% y bayes ingenuo 73.87%; para obtener estos resultados los autores hacen uso de distintos datos para entrenar los algoritmos, llegando a la conclusión que el algoritmo j48 es el mejor para su modelo predictivo ya que tiene mayor exactitud y un error mínimo en la fase de entrenamiento. En base a esto, la investigación nos muestra que debemos tener un buen conjunto de datos de entrenamiento y normalizarlos, para que en la fase de aprendizaje, con el uso del modelo de clasificación de machine learning, tenga un alto porcentaje de exactitud y un error mínimo en predicción.

(Kim A. & Lo A., 2010) Modelos de riesgo de crédito al consumidor a través de algoritmos de machine learning. Cambrigde, Estados Unidos. En esta investigación se pretendió realizar un modelo que efectúe la clasificación de riesgos de créditos con machine learning, para lo cual hacen uso de la técnica de árbol de decisiones siguiendo los pasos de identificación de las fuentes de datos, clasificación de entradas de modelos, comparación de métricas del modelo con CScore y finalmente la evaluación del modelo realizado. Con los pasos mencionados se desarrollan los algoritmos de árbol de decisiones y máquinas de soporte de vectores (SVM), los cuales ayudan en las fases de entrenamiento y mejora del modelo predictivo para lograr una mejor clasificación de los riesgos de crédito. Los resultados luego de aplicar ambos algoritmos están enfocados a la exactitud de predicción, el uso del algoritmo de árbol de decisiones con apoyo del algoritmo SVM logran una exactitud de clasificación de 85.00% aproximadamente lo que permite a esta investigación brindar un ahorro de costos de entre 6% y 8% sobre las pérdidas totales en cuanto a la asignación de créditos a los consumidores. Estos datos contribuyen para la clasificación de préstamos a través del uso de algoritmos de machine learning, puesto que permite un alto nivel de exactitud en predicción de nuevos datos; además, apoya a seguir una serie de pasos definidos para obtener un modelo matemático que pueda realizar un buen rendimiento haciendo validación de las métricas que

tenga el modelo a realizar independientemente del lenguaje en el que se desarrolle el algoritmo de aprendizaje automático.

(Kavitha K, 2016) Agrupación de solicitantes de préstamos basados en el porcentaje de riesgo utilizando la técnica de agrupamiento K-means. Kodaikanal, India. En esta investigación se pretende solucionar el problema del comportamiento de solicitantes de préstamos y la predicción de solicitudes haciendo uso de minería de datos y algoritmos basados en machine learning, debido a que los préstamos juegan un papel importante, se trata de analizar la información y clasificarla, por lo cual basándose en técnicas de agrupamiento como lo es el algoritmo de K-means, se clasifican los antecedentes financieros de los clientes. Además los autores siguen un flujo para el desarrollo del sistema que consiste en procesar la base de datos, generar reglas de asociación, evaluar el riesgo, aplicar el algoritmo K-means de clasificación y finalmente realizar una regla de predicción; esto ayuda en el desarrollo del algoritmo para agrupar la información de manera correcta. Los resultados arrojan que el algoritmo K-means crea tres agrupadores que son Bajo (Low), Medio (Medium) y Alto (High) indicando el nivel de riesgo de realizar un préstamo, cabe resaltar que se logra una precisión de 80.00% sobre los resultados. Este hecho contribuye a la investigación en que se deben seguir fases definidas antes de realizar el algoritmo de clasificación y aprendizaje, ya que si se tiene bien definida la información de clasificación, se puede lograr mayor exactitud, por lo cual este es un punto importante a considerar en el desarrollo del modelo predictivo en la investigación.

(Malca S., 2015) Modelo algorítmico para la clasificación de una hoja de planta en base a sus características de forma y textura. Lima, Perú. En esta investigación se pretende brindar un modelo de clasificación de hojas con uso de minería de datos y aprendizaje automático, debido a que existe un gran volumen de familias y clases de plantas en el ecosistema y otras que van apareciendo, por eso se plantea la solución de aprendizaje automático para tener un inventario actualizado de nuevas especies encontradas. Se hace uso de minería de datos para la recolección de información sobre las familias y clases de las hojas, en posterior por medio de los algoritmos de aprendizaje como árboles de decisión, redes neuronales, redes bayesianas y K- vecino más cercano se construye el algoritmo para clasificación automática. Los resultados usando la herramienta Weka con los algoritmos mencionados indican que con el uso de redes neuronales tomando como parámetros la forma, color y textura de la hoja se obtiene una precisión de 93.95%, con el uso de máquina de vector de soporte (SVM) se obtiene una eficiencia de 92.00% sobre los datos. Estos hechos contribuyen a la investigación, ya que nos indica que se debe recopilar información de entrenamiento para el algoritmo de aprendizaje y esto se puede realizar por medio de minería de datos y luego verificar la eficiencia del algoritmo de aprendizaje para ver si el modelo realizado es el correcto con lo que se quiere realizar.

(Quispesaravia R. & Pérez W, 2015) Herramienta de análisis y clasificación de complejidad de texto en español. En esta investigación se pretende crear una herramienta web para clasificar textos con uso de aprendizaje automático debido al bajo nivel en comprensión lectora en adolescentes, ya que las lecturas no son las adecuadas para ellos. Por medio de algoritmos como regresión gaussiana, máquina de vectores de apoyo (SVM) y árbol de decisión se realizan los algoritmos de aprendizaje y se evalúan en la herramienta Weka para elegir el mejor algoritmo. Los resultados indicaron que el uso de algoritmo FilteredClassifier obtiene una precisión de 72% y el algoritmo OnerR21 obtiene una precisión menor a la de FilteredClassifier. Este hecho ayuda al desarrollo de la investigación a evaluar bien los modelos antes de implementarlos, para luego hacer el aprendizaje automático del mismo.

2.2. Bases teóricas

2.2.1. Inteligencia Artificial

2.2.1.1. Definición

El concepto de inteligencia artificial se refiere no solamente a la capacidad que tienen las máquinas para intentar comprender como los humanos piensan, sino que también se esfuerza en construir entidades inteligentes. La inteligencia artificial es una de las ciencias más recientes que abarca una variedad de sub-campos, que van desde áreas de propósito general, como el aprendizaje y la percepción, a otras más específicas como demostración de teoremas matemáticos, procesamiento de lenguaje natural, diagnóstico de enfermedades, entre otras. La inteligencia artificial sinteriza y automatiza tareas intelectuales y es, por lo tanto, relevante para cualquier ámbito de la actividad intelectual humana (Russell y Norving, 2004).

2.2.1.2. Importancia de la Inteligencia Artificial

La inteligencia artificial es un conjunto de herramientas que permite el avance de muchas investigaciones como el machine learning, el reconocimiento de imágenes, entre otras; pero además envuelve el uso de mucha información para el correcto funcionamiento de sistemas que emplean estas técnicas de inteligencia artificial (Konovalenko, 2013). Según Enrique Dans en su publicación “La medida de la importancia de la inteligencia artificial”, nos dice: Las verdaderas claves del futuro son el machine learning y la inteligencia artificial, ya que cada vez más son usadas prácticamente en todos los negocios e industrias, la inteligencia artificial está en el estado del arte pues ha tomado el control de muchas compañías para sacar ventajas competitivas sobre los competidores cercanos, la innumerable cantidad de datos con la que cuentan las compañías es un factor fundamental, la importancia radica en ver la cosas

que se tienen alrededor y sacar un máximo provecho, es pues la inteligencia artificial y sus múltiples áreas de investigación lo que hacen de este campo una nueva forma de enfocar las cosas para un mejor desarrollo.

2.2.1.3. Técnicas aplicadas a la Inteligencia Artificial

La inteligencia artificial es un campo amplio, por cual se dividen en algunos campos de investigación. Los siguientes campos de investigación son los que menciona el Instituto de Investigación en Inteligencia Artificial (CSIC), en su publicación denominada la Inteligencia artificial:

a) Resolución de problemas y búsqueda

La inteligencia artificial tiene como objetivo resolver problemas de índole diferente. Para poder cumplir objetivos, dado un problema es necesario formalizarlo para poder resolver. En este campo es donde se centran los conocimientos para formalizarlos y buscar formas de resolución. (Torra, 2011).

b) Sistemas basados en conocimiento

Los programas de inteligencia artificial necesitan incorporar conocimientos del dominio de la aplicación, por ejemplo, en la medicina, para resolver estos problemas la inteligencia artificial se centra en estos temas (Torra, 2011).

c) Inteligencia artificial distribuida

Durante los primeros años la inteligencia artificial era monolítica. Según Torra nos dice que con el avance de la tecnología los ordenadores multiprocesador e Internet, hay interés en soluciones distribuidas. Estas van desde versiones paralelas de métodos ya existentes a nuevos problemas relacionados con agentes autónomos – programas software con autonomía para tomar decisiones e interaccionar con otro – y existen cuatro temas relacionados a este campo que son: El lenguaje natural, la visión artificial, la robótica y el reconocimiento del habla.

d) Aprendizaje automático

El rendimiento de un programa puede incrementarse si el programa aprende de la actividad realizada y de sus propios errores, debe existir un medio por el cual se tenga control de las fases de entrenamiento y desarrollo de estos modelos. Existen también herramientas que permiten extraer conocimiento a partir de base de datos (Torra 2011).

2.2.2. Machine Learning

2.2.2.1. Historia del Machine Learning

En el artículo A short history of Machine learning, Bernard Marr nos indica los acontecimientos más importantes, que dieron origen a Machine Learning:

- En 1950, Alan Turing creó el “Turing Test” para determinar si una computadora es realmente inteligente.
- En 1952, Arthur Samuel escribió el primer programa de machine learning, El programa fue el juego de las damas.
- En 1957, Frank Rosenblatt diseño la primera red neuronal para computadoras, el perceptrón.
- En 1967, el algoritmo vecino cercano fue escrito, este algoritmo permitió a las computadoras usar un reconocimiento de patrones muy básico.
- En 1979, Los estudiantes de la Universidad de Stanford inventaron el “Stanford Cart”, que puede navegar por obstáculos en una habitación por su cuenta.
- En 1981, Gerald Dejong introduce el concepto de Aprendizaje Basado en la Explicación, en el que un ordenador analiza datos de entrenamiento y crea una regla general que puede seguir descartando datos sin importancia.
- En 1985, Terry Sejnowski inventó NetTalk, que aprende a pronunciar palabras igual que un bebé.
- En 1990, el trabajo sobre el aprendizaje automático cambia de un enfoque basado en el conocimiento a un enfoque basado en datos. Los científicos comienzan a crear programas para computadoras para analizar grandes cantidades de datos y sacar conclusiones - o "aprender" - de los resultados.
- En 1997, IBM’s Deep Blue supera al campeón del mundo en ajedrez.
- En 2006, Geoffrey Hinton asignó el termino aprendizaje profundo para explicar nuevos algoritmos que permiten a los ordenadores ver y distinguir objetos y texto en imágenes y videos.

2.2.2.2. Definición

“Machine learning es la programación de computadoras para optimizar un criterio de rendimiento utilizando datos de ejemplos o experiencias pasadas.” (Alpaydin, 2010, p. 42). Esta técnica crea sistemas que aprenden automáticamente; es decir, identifican patrones complejos en millones de datos

y se mejoran de forma autónoma con el tiempo para generar decisiones y resultados fiables. (Gonzáles, 2014).

2.2.2.3. Importancia del Machine Learning

Las empresas están generando constantemente, en forma exponencial, gran cantidad de datos. Utilizar esa información, aplicando Machine Learning correctamente es una gran ventaja competitiva, debido a que se puede obtener predicciones de alto valor para tomar mejores decisiones y realizar acciones de negocios (Gonzáles, 2014).

Machine Learning hoy en día ya no es como se veía en el pasado. Nació a raíz del reconocimiento de patrones y de la teoría de que los ordenadores pueden aprender sin ser programados para realizar tareas específicas, en aquel tiempo los investigadores interesados en la inteligencia artificial querían ver si los ordenadores podrían aprender de los datos. El aspecto iterativo de machine learning es importante porque como modelos están expuestos a nuevos datos que son capaces de adaptarse de forma independiente. Los modelos de machine learning aprenden de los cálculos anteriores para producir decisiones y resultados repetibles con un nivel de confianza muy elevado. Mientras que muchos algoritmos de machine learning han existido desde hace mucho tiempo, la capacidad de aplicar automáticamente los cálculos matemáticos complejos para grandes volúmenes de datos es un desarrollo reciente (Statistical Analysis System, 2016).

El resurgir del interés por temas de machine learning – aprendizaje automático – se debe a los mismos factores que han hecho de la minería de datos y análisis bayesiano más popular hoy en día. Cosas como los volúmenes y variedades de los datos disponibles, el procesamiento computacional que es más barato y más potente, y el almacenamiento de datos accesible en crecimiento (Statistical Analysis System, 2016).

2.2.2.4. Minería de Datos y Machine Learning

La minería de datos se define como el proceso de descubrir patrones en los datos. El proceso debe ser automático o (más habitualmente) semiautomático. Los patrones descubiertos deben ser significativos en el sentido de que conducen a alguna ventaja, usualmente una ventaja económica. Los datos están invariablemente presentes en cantidades sustanciales. El objetivo es crear un proceso automatizado que toma como punto de partida los datos y cuya meta es la ayuda a la toma de decisiones. Gonzáles (2014), nos indica que una vez identificados los patrones, se pueden hacer predicciones con

nuevos datos que se incorporen al sistema. Por ejemplo los datos históricos de las compras de libros en una web online se pueden usar para analizar el comportamiento de los clientes en sus procesos de compra (títulos visitados, categorías, historial de compras…), agruparlos en patrones de comportamiento y hacer recomendaciones de compra a los clientes nuevos que siguen los patrones ya conocidos o aprendidos.

2.2.2.5. Algoritmos de Machine Learning

¿Qué algoritmo de machine learning se debe utilizar? Depende del tamaño, la calidad y la naturaleza de los datos, depende de qué se desea hacer con la respuesta. Y también depende del tiempo que disponga. (Microsoft, 2016).

a) Regresión lineal

La regresión lineal ajusta una línea (plano o hiperplano) al conjunto de datos; es potente, simple y rápida. (Microsoft, 2016). La regresión lineal pertenece a la categoría de aprendizaje supervisado, y es de regresión ya que su salida son valores continuos. El objetivo de la regresión es minimizar el error entre la función aproximada y el valor de la aproximación. Whitten y Frank (2005), nos indican que cuando el resultado o la clase es numérico y todos los atributos son numéricos, la regresión lineal es una técnica natural a considerar. Además, la idea es expresar la clase como una combinación de los atributos, con pesos determinados:

$$x = w_0 + w_1 a_1 + w_2 a_2 + \cdots + w_k a_k$$

Donde x es la clase; $a_1, a_2, \ldots, a_k$ son los valores de los atributos; y $w_0, w_1, \ldots, w_k$ son los pesos. Los pesos se calculan a partir de los datos de entrenamiento. Es necesario una forma de expresar los valores de entrenamiento.

b) Regresión logística

La regresión logística es en realidad una herramienta eficaz para la clasificación de dos clases y multiclase, es rápida y sencilla. Al utilizar una curva con forma de S en lugar de una línea recta la hace ideal para dividir los datos en grupos (Microsoft, 2016). Whitten y Frank (2005) indican que la regresión logística construye un modelo lineal basado en un variable objetivo. Supongamos primero que sólo hay dos clases. La regresión logística reemplaza la variable objetivo original

$$\Pr[1|a_1, a_2, \ldots, a_k],$$

Que no se puede aproximar con precisión usando una función lineal, con

$$\log(\Pr[1|a_1, a_2, \ldots, a_k]) / (1 - \Pr[1|a_1, a_2, \ldots, a_k]$$

Los valores resultantes ya no están restringidos al intervalo de 0 a 1, pero pueden estar en cualquier lugar entre el infinito negativo y el infinito positivo, esta función que define estos valores se llama función sigmoidal; además se apoya en el uso del método de optimización de la gradiente de descenso. La variable transformada se aproxima usando una función lineal igual que las generadas por regresión lineal. El modelo resultante es con pesos w.

$$\Pr[1|a_1, a_2, \ldots, a_k] = 1/ (1 + \exp(-w_0 - w_1 a_1 - \cdots - w_k a_k)),$$

c) K-means

K-means es un algoritmo de aprendizaje sin supervisión. Define un prototipo en términos de un centroide, que suele ser la medida de un grupo de puntos, y se aplica típicamente a objetos en un espacio n-dimensional continuo (Tan, Steinbach, Kumar, p. 496, 2006).

Consiste en dividir N observaciones en k grupos, donde cada elemento se asigna al grupo cuya media le sea más cercana. Esta técnica de agrupamiento es simple; en primer lugar, elegir K centroides inicial, es decir, el número de clusters deseados. A continuación, cada punto se asigna al centroide cercano y cada grupo se actualiza en función de los puntos asignados. Se repiten los puntos de asignación y actualización hasta que ningún punto cambie de clúster, o hasta que los centroides permanezcan igual (Tan, Steinbach, Kumar, p. 497, 2006).

Algorithm 1.5 K-Means

Cluster($\mathbf{X}$) {Cluster dataset $\mathbf{X}$}
  Initialize cluster centers μ_j for $j = 1, \ldots, k$ randomly
  repeat
    for $i = 1$ **to** m **do**
      Compute $j' = \text{argmin}_{j=1,\ldots,k} d(x_i, \mu_j)$
      Set $r_{ij'} = 1$ and $r_{ij} = 0$ for all $j' \neq j$
    end for
    for $j = 1$ **to** k **do**
      Compute $\mu_j = \frac{\sum_i r_{ij} x_i}{\sum_i r_{ij}}$
    end for
  until Cluster assignments r_{ij} are unchanged
  return $\{\mu_1, \ldots, \mu_k\}$ and r_{ij}

Figura Nº 1 Algoritmo K-Means
Fuente: Smola y Vishwanathan. Introduction to Machine Learning.

d) Máquinas de Soporte Vectorial (SVM)

Las máquinas de soporte vectorial buscan el límite que separa las clases con el mayor margen posible, cuando no se puede separar bien las dos clases, los algoritmos buscan el mejor límite que pueden. (Microsoft, 2016).

Una SVM primero mapea los puntos de entrada a un espacio de características de una dimensión mayor y encuentra un híper plano que los separa y maximice el margen m entre las clases en este espacio como se aprecia en la Figura 2 (Betancourt, 2005, pg. 67).

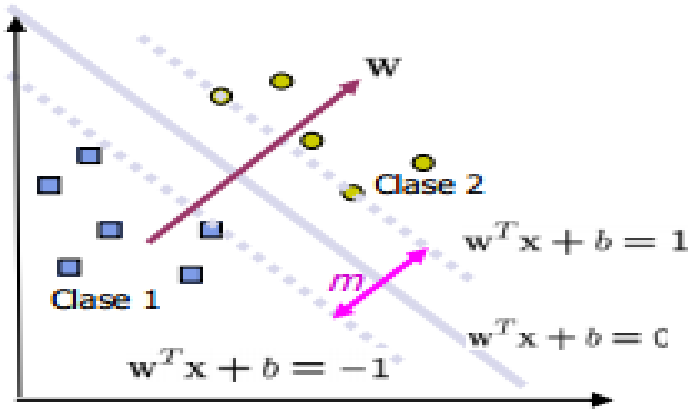


Figura Nº 2 Frontera de decisión para SVM

Fuente: Betancourt. Las máquinas de soporte vectorial

e) Naive Bayes

El clasificador Naive Bayes es uno de los clasificadores de las redes bayesianas más eficiente en el proceso de clasificación. Sus principales cualidades son su simplicidad y precisión, aunque su estructura siempre es fija se ha demostrado que tiene una alta precisión de clasificación y un error mínimo. (Como lo citó Barrientos y otros, 2008, p.34)

En la Figura 3 se puede apreciar la estructura de este tipo de clasificador, en la que existe un nodo para las variables de clase, esta es el padre de todas las demás variables. Además, no se permiten arcos entre las variables, asumiendo la restricción de que los atributos son independientes conocido el valor de la variable clase.

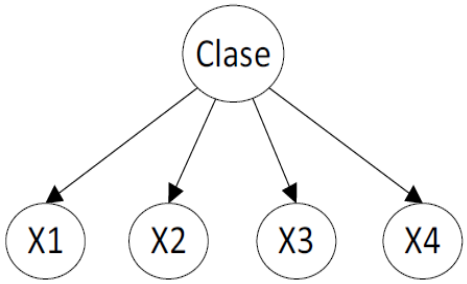


Figura Nº 3 Estructura de un clasificador Naive Bayes

Fuente: Barrientos Martínez, R., Cruz Ramírez, N., Acosta Meza, H., Rabatte Suarez, I., otros. Evaluación del Potencial de Redes Bayesianas en la Clasificación en Datos Médicos.

f) Redes neuronales artificiales

Las redes neuronales son algoritmos de aprendizaje inspirados en el cerebro que abarcan problemas multiclase, de dos clases y de regresión. Las redes neuronales pueden tardar mucho tiempo para entrenarse, especialmente para grandes conjuntos de datos con muchas características; también tiene más parámetros que la mayoría de los algoritmos (Microsoft, 2005). Cada neurona se representa como una unidad de proceso que forma parte de una entidad mayor, la red neuronal.

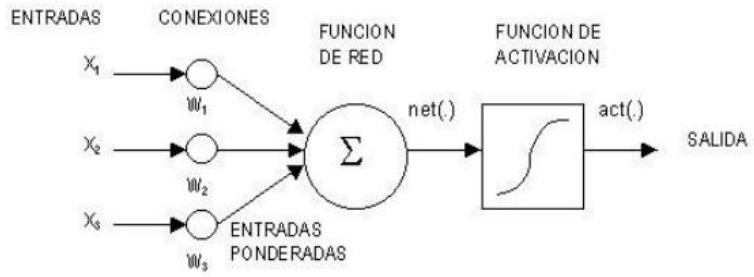


Figura Nº 4 Neurona artificial

Fuente: Sánchez Morales. Máquinas de aprendizaje extremo multicapa: Estudio y evaluación de resultados en la segmentación automática de carótidas en imágenes ecográficas

2.2.2.6. Modelos de aprendizaje

2.2.2.6.1. Aprendizaje supervisado

En este tipo de aprendizaje se le asigna al algoritmo un conjunto del que se sabe cómo va ser la salida, existe una relación entre la entrada y la salida. El objetivo es aprender un mapeo de la entrada a una salida cuyos valores correctos son proporcionados por un supervisor (Alpaydin, 2010, pg. 11). Kevin (s/a), nos explica que en el aprendizaje supervisado la finalidad es aprender un mapeo de las

entradas “x” a las salidas “y”, dado un conjunto etiquetado de pares de entrada y salida.

$$D = \{[x_i, y_i]\}_{i=1}^{N}$$

Aquí D es el conjunto de entrenamiento y N, el número de ejemplos de entrenamiento. Cada entrada de entrenamiento x_i es un vector D-dimensional de números, que representa, digamos, la altura y el peso de una persona. Estos se llaman características, atributos o covariables. Del mismo modo, la forma de la variable de salida o respuesta puede en principio ser cualquier cosa, pero la mayoría de los métodos asumen que y_i es una variable categórica o nominal de algún conjunto finito, $y_i \in \{1, \dots, C\}$, o que y_i es un escalar real-valorado (tal como nivel de renta). Cuando y_i es categórico, el problema se conoce como clasificación o reconocimiento de patrones, y cuando y_i es valor real, el problema se conoce como regresión. Otra variante, conocida como regresión ordinal, ocurre donde el espacio de etiqueta “y” tiene algún ordenamiento natural, como los grados A-F.

Clasificación: El objetivo es aprender un mapeo de entradas “x” a salidas “y”, donde “y” ∈ {1,..., C}, siendo C el número de clases. Si C = 2, esto se llama clasificación binaria; Si C> 2, se denomina clasificación multiclass. Si las etiquetas de la clase no son mutuamente excluyentes (por ejemplo, alguien puede ser clasificado como alto y fuerte), se llama clasificación multietiqueta, pero esto se ve mejor como predicción de múltiples etiquetas de clase binarias relacionadas (un modelo de salida múltiple) (Kevin, s/a, pg. 3).

Regresión: La variable de respuesta es continua. Se tiene una única entrada de valor real xi ∈ R, y una única respuesta real yi ∈ R. Se considera la posibilidad de ajustar dos modelos a los datos: una recta y una función cuadrática. Pueden surgir varias extensiones de este problema básico, tales como tener entradas de alta dimensionalidad, valores atípicos, respuestas no lisas, etc. (Kevin, s/a, pg. 8).

2.2.2.6.2. Aprendizaje no supervisado

En el aprendizaje no supervisado el proceso se lleva a cabo teniendo información del conjunto de datos formados tan solo por la entrada, no se tiene información sobre el espacio de salida. Este modelo tiene como objetivo encontrar las regularidades en la entrada. Hay una

estructura al espacio de entrada tal que ciertos patrones ocurren más a menudo que otros, y queremos ver qué sucede generalmente y qué no. (Alpaydin, 2010, pg. 11). Kevin (s/a), nos explica que en el aprendizaje no supervisado solo se nos da entradas,

$$D = \{[x_i]\}_{i=1}^{N}$$

Y el objetivo es encontrar patrones interesantes en los datos; es decir, descubriendo conocimiento. No se indica el tipo de patrones, y no hay métrica de error obvio para usar. Como ejemplo se considera el problema de agrupar datos en grupos. Por ejemplo, se traza algunos datos 2d, que representan la altura y el peso de un grupo de 210 personas. Parece que podría haber varios grupos, o subgrupos, aunque no están claros cuántos. Sea K el número de grupos. Nuestro primer objetivo es estimar la distribución sobre el número de grupos, p (K | D); Esto nos dice si hay subpoblaciones dentro de los datos. Por simplicidad, a menudo se aproxima la distribución p (K | D) por su modo, K * = arg maxK p (K | D). En el caso supervisado, nos dijeron que hay dos clases (masculina y femenina), pero en el caso sin supervisión, se puede elegir tantos o pocos grupos como se desea.

Estas son algunas aplicaciones del mundo real de la agrupación.

- En astronomía, el sistema de auto clase (Cheeseman et al., 1988) descubrió un nuevo tipo de estrella, basado en el agrupamiento de mediciones astrofísicas.
- En el comercio electrónico, es común agrupar a los usuarios en grupos, basándose en su comportamiento de compra o de navegación por Internet, y luego enviar publicidad personalizada dirigida a cada grupo.
- En biología, es común agrupar datos de citometría de flujo en grupos, para descubrir diferentes subpoblaciones de células.

2.2.3. Préstamo crediticio

2.2.3.1. Definición

Un préstamo crediticio es una operación por la cual una entidad financiera pone a disposición del cliente una cantidad de dinero determinada que se estipula por medio de un contrato, en el que se adquiere la obligación de devolver ese dinero en un tiempo establecido. En el préstamo, una de las partes – el prestamista – es el que pone todas las condiciones para que sea factible la

entrega del monto solicitado, además esto incluye la devolución con los intereses pactados en uno o varios pagos programados (BBVA, 2016).

2.2.3.2. Importancia en la economía

Los préstamos crediticios o préstamos bancarios no solo se limitan al proceso mismo del préstamo, cuando el acceso al financiamiento es limitado, también se restringen las posibilidades de crecimiento de una economía. En los países en desarrollo una gran parte de la población no tiene acceso al crédito, cuando no hay acceso al crédito el consumo de las familias y la inversión de las empresas debe financiarse con los ingresos de cada periodo. Esto puede generar inconvenientes cuando los ingresos son muy variables. El uso responsable del crédito facilita realizar gastos de consumo e inversión por encima de lo que permiten los ingresos corrientes. Sin embargo, el monto de financiamiento que una familia, una empresa o una economía recibe siempre está asociado a su capacidad de pagar sus deudas. En la economía peruana existe aún mucho margen para el crecimiento del crédito, el ratio de crédito a PBI es bajo en comparación con otros países de la región. (BCR, 2009). Los préstamos son parte vital de una entidad financiera, la gran mayoría de negocios tienen intervención de estas entidades ya que siempre necesitan conseguir un préstamo. Los bancos hacen dinero tomando los fondos de los depositantes y de otras fuentes y con esto prestan dinero a los clientes (Vindi N, 2012).

2.2.3.3. Proceso de evaluación de un préstamo crediticio

Tanto para personas naturales como para empresas es necesario seguir ciertos criterios de evaluación para acceder a un préstamo crediticio. No es necesario ser cliente de una entidad financiera, pero sí se necesita tener una buena clasificación en el sistema financiero, es decir, no reportar atrasos en pagos de deudas en la central de riesgos de la Superintendencia de Banca y Seguros, asimismo, la persona que desea solicitar el préstamo debe sustentar ingresos netos desde mil soles o su equivalente en dólares y tener continuidad laboral mínima de seis meses. Dependiendo la entidad financiera donde se solicite el préstamo crediticio la presentación de documentación puede variar, pero algunos de estos documentos son: documento nacional de identidad (DNI) del solicitante o cónyuge; recibo de luz, agua, o telefonía fija; copia del registro único del contribuyente (RUC), declaración jurada anual de impuestos a la renta; tres últimos PDT y/o tres últimos recibos por honorarios profesionales; dos últimas declaraciones juradas anuales de impuesto a la renta; última boleta de pago de ingresos fijos y dos últimas boletas de pago por ingresos variables;

contrato de locación vigente. También existen algunas restricciones a tener en cuenta para aplicar a un préstamo, una persona no puede ser mayor de 70 o 75 años (BBVA, 2017).

2.2.3.4. Riesgos de un préstamo crediticio

Un préstamo crediticio, a pesar de todos los beneficios que reporta, también cuenta con una serie de riesgos de los que se debe tener en cuenta antes de solicitarlo. Algunos de estos riesgos que se debe tener en cuenta son:

- Conocer la capacidad de endeudamiento para hacer frente a las cuotas que exija el crédito de manera mensual.
- Utilizar el dinero del crédito de manera inteligente, usarlo de manera adecuada para no endeudarnos sin necesidad.
- Evitar tener abierto varios créditos, ya que, a mayor número de crédito mayor compromiso crediticio y mayor dificultad para asumirlo.
- Conocer bien el préstamo crediticio que se está adquiriendo, para así evitar situaciones inesperadas como recargos o intereses excesivos.

El riesgo de un préstamo puede ocasionar endeudamiento excesivo por parte de las personas que piden estos préstamos, pudiendo llevar a las entidades bancarias a embargar bienes que no se acuerdan en un inicio. Es importante considerar todas las posibilidades. Por parte de los bancos el riesgo de un préstamo crediticio está en base a las condiciones que tiene la persona para pagar dicho préstamo en un tiempo establecido es por eso que tienen que hacer un análisis muy detallado de la situación en la que se va a conceder un préstamo (Llorca, 2012).

2.2.3.5. Tecnología y préstamos crediticios

La tecnología está en las entidades bancarias desde hace mucho tiempo, y sobre todo el uso de machine learning en estas entidades, es una de las primeras en aplicarse.

En algunas entidades crediticias la concesión de un préstamo depende más de la educación o la personalidad de un cliente, que de su calificación financiera. Existen varias startups con soluciones que permiten disminuir, gracias a algoritmos, los niveles de morosidad de bancos y empresas de concesión de préstamos o prestar dinero con menos riesgos con otros modelos de evaluación (BBVAOPEN4U, 2015).

Algunas de estas startups son:

a) Earnest

Earnest utiliza tecnología y datos para alejarse del modelo tradicional de concesión de préstamos en EEUU: se puede acceder a créditos en buenas condiciones si tu ratio crediticio es bueno. En earnest el funcionamiento es distinto, no sólo es importante una buena calificación financiera, también lo son aspectos como la educación, el historial laboral o cuál es si situación financiera en el momento de solicitar préstamo. Earnest crea un perfil individual de cada solicitante mediante el uso de algoritmos y análisis predictivo que evalúan responsabilidad y potencial.

b) Zestfinance

Zestfinance fue fundada por Douglas Merril, vicepresidente de Ingeniería de Google, grupo conformado por matemáticos e ingenieros en computación. Zestfinance está basado en el campo de machine learning – aprendizaje automático – para aplicar los datos a los créditos de forma eficiente.

Zestfinance evalúa numerosas variables del usuario para reducir el riesgo de fraude o impago y establecer una relación comercial con el cliente a largo plazo. La compañía presume de una mejora de ese ratio del 40%, lo que permite mayor disponibilidad de crédito, mejores intereses y seguridad de cobro para los prestamistas o entidades bancarias que son sus clientes finales. Todo esto se consigue con algoritmos de machine learning y análisis predictivo.

Métodos como el aprendizaje automático – machine learning – o el aprendizaje profundo – deep learning – están ayudando a las entidades financieras en numerosos campos operativos, y lógicamente, las APIs especializadas en machine learning y deep learning son el punto de partida para cualquier transformación. Gracias a ellas los bancos pueden crear productos finalistas que aporten valor a la entidad y a sus clientes, permiten extraer datos relevantes de Big Data, búsqueda de patrones que faciliten ofertas más personalizadas, ajustes de precios o detección de procesos de fraude bancario (OPENBBVA4U, 2016).

2.2.4. Metodología

2.2.4.1. Extreme Programming

2.2.4.1.1. Definición

Extreme Programming fue introducida como metodología ágil de desarrollo de software sobre finales de los 90’s, surge como una

nueva manera de encarar proyectos de software, proponiendo una metodología basada esencialmente en la simplicidad y agilidad. Las metodologías de desarrollo de software tradicionales (ciclo de vida en cascada, evolutivo, en espiral, iterativo, etc.) aparecen, comparados con los nuevos métodos propuestos en XP, como pesados y poco eficientes. La metodología XP está diseñada para entregar el software que los clientes necesitan en el momento en que lo requieran. Extreme Programing alienta a los desarrolladores a responder a los requerimientos cambiante de los clientes, aún en fases tardías del ciclo de vida del desarrollo (Fowler, 2005).

2.2.4.1.2. Ciclo de vida de software XP

La metodología XP define cuatro variables para cualquier proyecto de software: costo, tiempo, calidad y alcance. Además, se especifica que, sólo tres de ellas podrán ser fijadas arbitrariamente por actores externos al grupo de desarrolladores.

El ciclo de vida de un proyecto XP incluye, al igual que las otras metodologías, entender lo que el cliente necesita, estimar el esfuerzo, crear la solución y entregar el producto final al cliente. Sin embargo, XP propone un ciclo de vida dinámico, donde se admite expresamente que, en muchos casos, los clientes no son capaces de especificar sus requerimientos al comienzo de un proyecto. Por eso se trata de realizar ciclos de desarrollo cortos, denominados iteraciones, con entregables funcionales al finalizar cada ciclo. En cada fase se realiza un ciclo completo de análisis, diseño, desarrollo y pruebas, pero utilizando un conjunto de reglas y prácticas que caracterizan a XP (Joskowics, 2005).

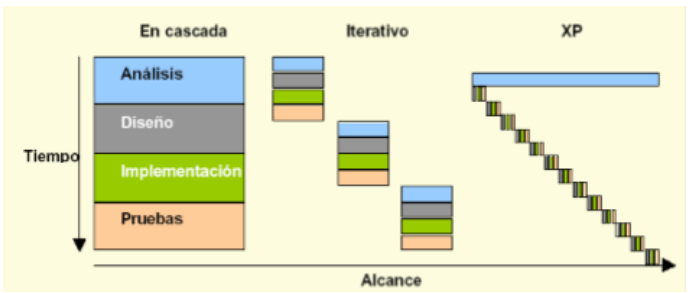


Figura Nº 5 Comparación en tiempo de las metodologías de desarrollo de software

Fuente: Cueva J. & Acebal C. Extreme Programming (XP). Un nuevo método de desarrollo de software

a) Fase de exploración

En esta fase se define el alcance general del proyecto, donde el cliente define lo que necesita mediante la redacción de sencillas historias de usuarios. Los programadores estiman los tiempos de desarrollo en base a esta información. Debe quedar claro que las estimaciones realizadas en esta fase son primarias, y podrían variar cuando se analicen más en detalle en cada iteración.

Esta fase dura típicamente un par de semanas, y el resultado es una visión general del sistema, y un plazo total estimado.

b) Fase de planificación

La planificación es una fase corta, en la que el cliente, los gerentes y el grupo de desarrolladores acuerdan el orden en que deberán implementarse las historias de usuario, y, asociadas a éstas, las entregas. Típicamente esta dase consiste en una o varias reuniones grupales de planificación. El resultado de esta fase es un Plan de Entregas, o “Release Plan”.

c) Fase de iteraciones

Esta es la fase principal en el ciclo de desarrollo de XP. Las funcionalidades son desarrolladas en esta fase, generando al final de cada una un entregable funcional que implementa las historias de usuario asignadas a la iteración. Como las historias de usuario no tienen suficiente detalle como para permitir su análisis y desarrollo, al principio de cada iteración se realizan las tareas necesarias de análisis, recabando con el cliente todos los datos que sean necesarios. El cliente, por lo tanto, también debe participar activamente durante esta fase del ciclo.

Las iteraciones son también utilizadas para medir el progreso del proyecto. Una iteración terminada sin errores es una medida clara de avance.

d) Fase de puesta en producción

Si bien al final de cada iteración se entregan módulos funcionales y sin errores, puede ser deseable por parte del

cliente no poner el sistema en producción hasta tanto no se tenga la funcionalidad completa.

En esta fase no se realizan más desarrollos funcionales, pero pueden ser necesarias tareas de ajuste.

2.2.4.1.3. Reglas y Prácticas

Según Well (2006), la metodología XP tiene un conjunto importante de reglas y prácticas. En forma genérica, se pueden agrupar en:

- Reglas y prácticas para la planificación
- Reglas y prácticas para el diseño
- Reglas y prácticas para el desarrollo
- Reglas y prácticas para las pruebas

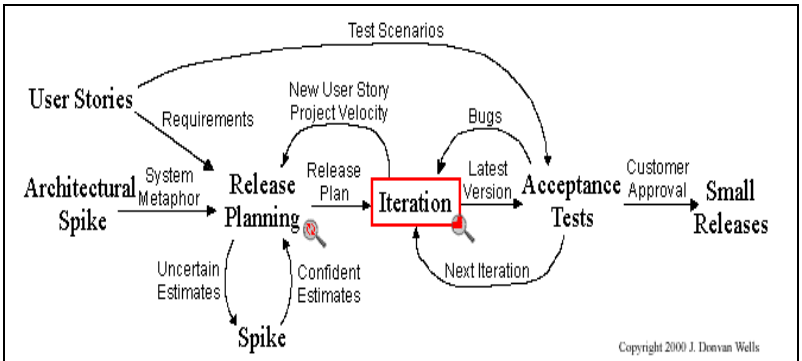


Figura Nº 6 Fases de metodología XP

Fuente: Extreme Programming Project, Fases de metodología XP

a) Planificación

El proyecto comienza recopilando historias de usuarios, las que sustituyen a los tradicionales casos de uso. Una vez obtenidas las historias de usuario, los programadores evalúan rápidamente el tiempo de desarrollo de cada una. Si alguna de ellas tiene riesgos que no permiten establecer con certeza la complejidad del desarrollo, se realizan pequeños programas de prueba (spikes), para reducir los riesgos, luego se realiza un cronograma de entregar (release plan) en los que todos estén de acuerdo. Una vez acordado este cronograma, comienza una fase de iteraciones, en dónde en cada una de ellas se desarrolla, prueba e instala unas pocas historias de usuarios.

Según Flower (2001), los planes XP se diferencian de las metodologías en tres aspectos que son:

- Simplicidad del plan
- Los planes se realizan por las personas que realizarán el trabajo
- Los planes no son predicciones del futuro

Los conceptos básicos de esta planificación son los siguientes:

- Historias de usuario
- Plan de entregas
- Plan de iteraciones
- Reuniones diarias de seguimiento

b) Diseño

La metodología XP hace especial énfasis en los diseños simples y claros. Los conceptos más importantes de diseño en esta metodología son los siguientes:

Simplicidad

Un diseño simple se implementa con más rapidez que uno complejo. Por ello XP propone implementar el diseño más simple posible que funcione. Se siguiere nunca adelantar la implementación de funcionalidades que no correspondan a la iteración en la que se esté trabajando.

Soluciones "spike"

Cuando aparecen problemas técnicos, o cuando es difícil de estimar el tiempo para implementar una historia de usuario, pueden utilizarse pequeños programas de prueba, llamados spike, para explorar diferentes soluciones.

c) Desarrollo

Uno de los requerimientos de XP es tener al cliente disponible durante todo el proyecto. Al comienzo del proyecto, el cliente debe proporcionar las historias de usuarios. Pero, dado que estas historias son expresamente cortas y de alto nivel, no contienen los detalles necesarios para realizar el desarrollo del

código. Estos detalles deben ser proporcionados por el cliente, y discutidos con los desarrolladores, durante la etapa de desarrollo.

Uso de estándares

Si bien esto no es una idea nueva, XP promueve la programación basada en estándares, de manera que sea fácilmente entendible por todo el equipo, y que facilite la recodificación.

Programación dirigida por las pruebas

Las pruebas a las que se refiere esta práctica, son las pruebas unitarias realizadas por los desarrolladores. La definición de estos test al comienzo, condiciona o dirige el desarrollo (Miller, 2003).

Programación en pares

XP propone que se desarrolle en pares de programadores, ambos trabajando juntos en un mismo ordenador, permite minimizar errores y se logran mejores diseños, compensando la inversión en horas. El producto obtenido es por lo general de mejor calidad que cuando el desarrollo se realiza por programadores individuales (Williams & Cockburn, 2000).

d) Pruebas

Las pruebas unitarias son una de las piedras angulares de XP. Todos los módulos deben de pasar las pruebas unitarias antes de ser liberados o publicados. Por otra parte, las pruebas deben ser definidas antes de realizar el código. Que todo código liberado pase correctamente las pruebas unitarias es lo que habilita que funcione la propiedad colectiva del código.

Detección y corrección de errores

Cuando se encuentra un error (bug), éste debe ser corregido inmediatamente, y se deben tener precauciones para que errores similares no vuelvan a ocurrir. Asimismo, se generan nuevas pruebas para verificar que el error haya sido resuelto.

Pruebas de aceptación

Las pruebas de aceptación son creadas en base a las historias de usuario, en cada ciclo de la iteración del desarrollo. El cliente debe especificar uno o diversos escenarios para comprobar que una historia de usuario ha sido correctamente implementada.

2.2.4.2. CRISP

2.2.4.2.1. Definición

Cross-Industry Standard Process for Data Mining es un método probado para orientar sus trabajos de minería de datos y machine learning. Es flexible y se puede personalizar fácilmente. En la práctica, muchas de las tareas se pueden realizar en un orden diferente y con frecuencia será necesario retroceder a las tareas anteriores y repetir ciertas acciones.

Según, IBM:

- Como metodología incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.
- Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.

2.2.4.2.2. Fases de la metodología CRISP

Este modelo contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases. La secuencia de las fases no es estricta.

Figura Nº 7 Ciclo vital del modelo
Fuente: IBM. Manual CRISP-DM de IBM SPSS Modeler

a) Comprensión del negocio

El objetivo de esta etapa del proceso es descubrir los factores importantes que podrían influir en el resultado del proyecto, esta fase permite entender los objetivos y requerimientos del proyecto desde una perspectiva del negocio.

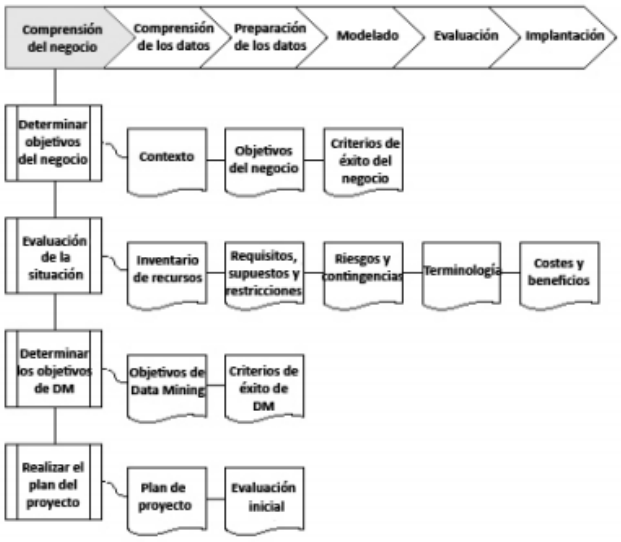


Figura Nº 8 Fase de Comprensión del negocio

Fuente: Galán Cortina, Víctor. Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario.

A continuación, se describen las tareas que componen esta fase:

- Determinar objetivos del negocio: Según IBM, en esta tarea se debe obtener la máxima información posible de los objetivos comerciales.
- Evaluación de la situación: Según IBM, después de definir los objetivos, es necesario realizar la evaluación de la situación actual.
- Determinar los objetivos: Los objetivos deben ser claros, para seguir eso se pueden realizar tareas que describan el tipo de problema como conglomerado, pronóstico o clasificación, documentar objetivos técnicos y proporcionar datos reales para resultados deseados (IBM, pg. 10).
- Realizar el plan de proyecto: En este punto, lo planteado y los objetivos formarán la base de este. La correcta escritura del

plan de proyecto permite informar a todos los usuarios relacionados con los objetivos, recurso, riesgos de proyecto y programar todas las fases (IBM, pg. 11).

b) Comprensión de los datos

La comprensión de datos implica acceder a los datos y explorarlos con la ayuda de tablas y gráficos. De esta forma se podrá determinar la calidad de los datos y describir los resultados de estos pasos en la documentación del proyecto (IBM, pg. 14).

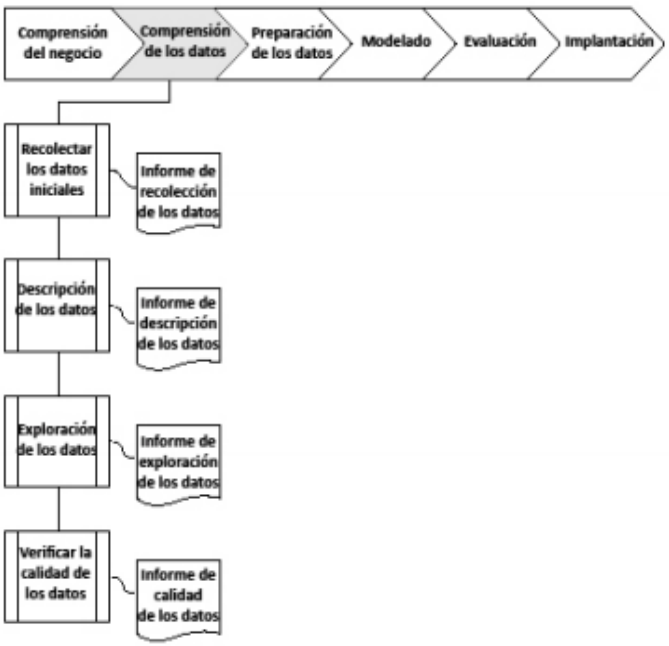


Figura Nº 9 Fase de Comprensión de los datos

Fuente: Galán Cortina, Víctor. Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario.

A continuación, se describen las tareas que propone esta fase:

- Recopilación de datos iniciales: En esta fase se indican las fuentes de donde provienen los datos que puede ser diferentes.
  - ✓ Datos existentes: Incluye datos transaccionales, datos de encuestas, registros Web, etc.
  - ✓ Datos adquiridos: Considerar si la organización utiliza datos adicionales.

- ✓ Datos adicionales: Realizar encuestas o seguimientos, en caso los datos anteriores no son suficientes.

- Descripción de los datos: La descripción de datos se centra en la cantidad y calidad de los datos. Se consideran las siguientes características para describir los datos (IBM, pg. 15):
  - ✓ Cantidad de datos: Incluir datos estadísticos de tamaños para todos los conjuntos de datos y tener en cuenta tanto el número de registros como los campos cuando describa los datos.
  - ✓ Tipos de valores: Prestar atención al tipo de valor para evitar posteriores problemas.
  - ✓ Esquemas de codificación: Los valores de la base de datos son representaciones de características como género o tipo de producto, registrar los esquemas incoherentes en el informe de datos.
- Exploración de los datos: Esta fase ayuda a formular la hipótesis y dar forma a las tareas de transformación de datos.
- Verificar la calidad de los datos: Es necesario realizar un análisis de la calidad de los datos antes de proceder al modelado, para evitar problemas.

c) Preparación de los datos

Es una fase muy importante, se estima que la preparación de datos suele llevar 50 – 70% del tiempo y esfuerzo del proyecto (IBM, pg. 21).

Figura Nº 10 Fase de Preparación de los datos

Fuente: Galán Cortina, Víctor. Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario.

La preparación de datos suele implicar las tareas siguientes:

- Fusionar conjuntos y/o registros de datos.
- Seleccionar de una muestra de un subconjunto de datos.
- Agregar registros.
- Derivar nuevos atributos.
- Clasificar los datos para el modelado.
- Eliminar o sustituir de valores en blanco o ausente.
- Dividir en conjuntos de datos de prueba y entrenamiento.
- Selección de datos: Dos formas de seleccionar datos:
  - ✓ Selección de elementos (filas).
  - ✓ Selección de atributos o características (columnas).

- Limpieza de datos: La limpieza de datos implica observar más de cerca los problemas en los datos que ha seleccionado incluir en el análisis.
- Construcción de nuevo datos: Existen dos formas de construir datos:
  - ✓ Derivación de atributos (columnas o características).
  - ✓ Generación de registros (filas).
- Integración de datos: Existen dos métodos básicos para integrar los datos:
  - ✓ La fusión de datos implica unir dos conjuntos de datos con registros similares, pero con atributos diferentes.
  - ✓ La adición de datos implica integrar dos o más conjuntos de datos con atributos similares, pero con registros diferentes.
- Formateo de los datos: Como paso final antes de la construcción del modelo, es muy útil comprobar si algunas técnicas requieren aplicar un formato concreto a la clasificación de los datos.

d) Modelado

En esta fase los datos que han preparado se incorporan a las herramientas analíticas y los resultados comenzarán a arrojar algo de luz al problema planteado (IBM, pg. 27).

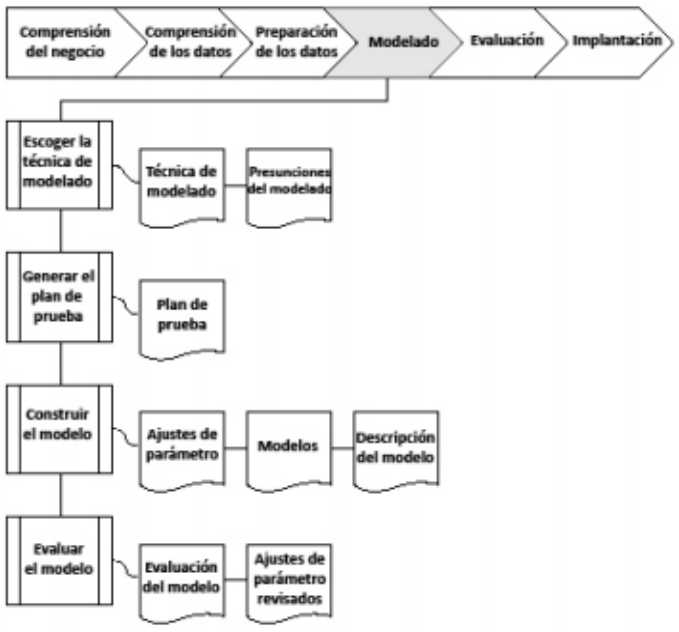


Figura Nº 11 Fase de Modelado

Fuente: Galán Cortina, Víctor. Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario.

A continuación, se describen las tareas que propone esta fase:

- Selección de técnicas de modelado: Según IBM, al decidir sobre el modelo a utilizar, se debe tener en cuenta los siguiente aspectos:
  - ✓ ¿Requiere el modelo que los datos se dividan en conjuntos de entrenamiento y prueba?
  - ✓ ¿Requiere el modelo un cierto nivel de calidad de datos? ¿Puede alcanzar este nivel con los datos que dispone?
  - ✓ ¿Son sus datos el tipo correcto para un modelo concreto? En caso contrario, ¿Puede realizar las conversiones necesarias utilizando nodos de manipulación de datos?
  - ✓ Modelado de supuestos: Documentar cualquier supuesto de datos y modificación realizada para cumplir los requisitos del modelo.
- Generación de un diseño de comprobación: Existen dos formas para generar un diseño:

- ✓ Descripción de los criterios de “bondad” de un modelo, para modelos supervisados las mediciones de bondad suelen calcular la tasa de error de un modelo concreto; y para modelos sin supervisión, las mediciones pueden incluir criterios como facilidad de interpretación, distribución o el tiempo de procesamiento.
- ✓ Definición de los datos en los que se comprobarán estos datos.

- Generación de los modelos: Al finalizar la generación de modelos, se dispondrá de tres tipos de información:
  - ✓ Configuración de parámetros.
  - ✓ Los modelos reales producidos.
  - ✓ Descripción de resultados de modelos.
- Evaluación del modelo:
  - ✓ Evaluación global del modelo: Para cada modelo que se va a considerar, es una buena idea crear un método de valoración basado en los criterios generados en su plan de pruebas. Una vez haya realizado la valoración, clasifique los modelos en función de criterios objetivos y subjetivos. (IBM, pg. 32).

e) Evaluación

Evaluar los resultados de sus esfuerzos utilizando los criterios de rendimiento comercial establecidos en el inicio del proyecto. Es la clave para asegurar que su organización pueda utilizar los resultados que ha obtenido.

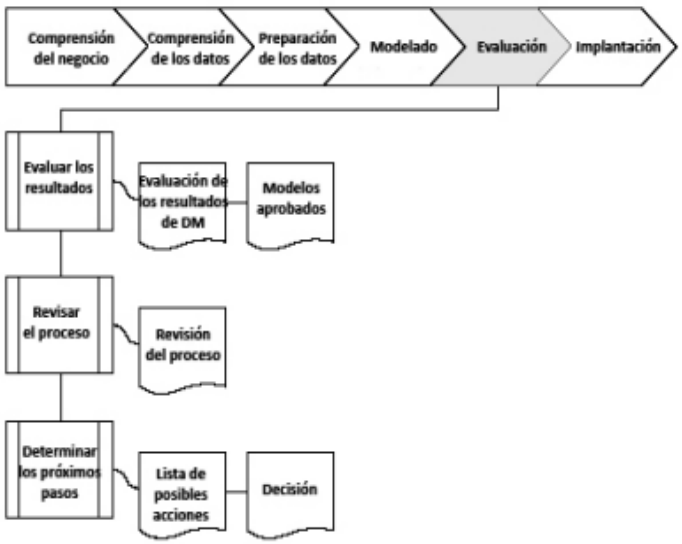


Figura Nº 12 Fase de Evaluación

Fuente: Galán Cortina, Víctor. Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario.

A continuación, se describen las tareas que propone esta fase:

- Evaluación de los resultados: Formalizar la evaluación en función de si los resultados del proyecto cumplen los criterios del rendimiento comercial.
- Proceso de revisión: Las metodologías eficaces suelen incluir tiempo para reflexionar sobre los aciertos y errores del proceso que se acaba de completar. Una parte fundamental de CRISP-DM es aprender de su propia experiencia para que sus proyectos de minería de datos sean más efectivos.
- Determinación de los pasos siguientes: Al llegar a esta fase se dispone de dos opciones.
  - ✓ Continuar con la fase de desarrollo: ayudará a incorporar los resultados del modelo a su proceso comercial y producir un informe final.
  - ✓ Volver y refinar o sustituir los modelos: Si encuentra que los resultados son casi óptimos, pero no lo suficiente, considere otro tipo de modelado.

f) Implantación

IBM, indica que este proceso consiste en utilizar sus nuevos conocimientos para implementar las mejoras en su organización. Además, la distribución puede significar que utilice los conocimientos adquiridos para aplicar modificaciones en su organización. La fase de distribución de CRISP-DM incluye dos tipos de actividades:

- ✓ Planificación y control de la distribución de los resultados.
- ✓ Finalización de tareas de presentación como la producción de un informe final y la revisión de un proyecto

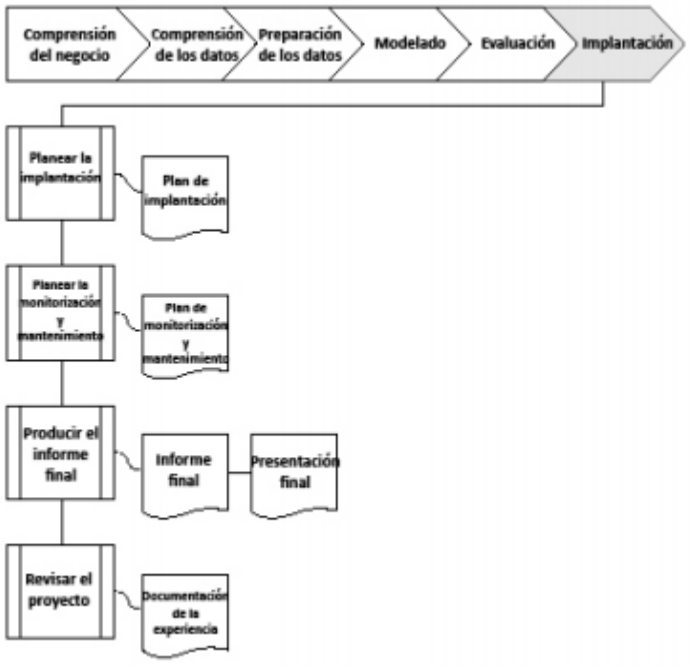


Figura Nº 13 Fase de Implantación

Fuente: Galán Cortina, Víctor. Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario.

A continuación, se describen las tareas que propone esta fase:

- Planificación de implantación: Para realizar una correcta distribución realizar las siguientes tareas:
  - ✓ Resumir los resultados; modelos y descubrimientos.
  - ✓ Crear una planificación paso a paso para la distribución e integración con sus sistemas.

- ✓ Crear un plan para difundir la información a los estrategas de la organización.
- ✓ ¿Dispone de planes de distribución alternativos para ambos tipos de resultados que se deben documentar?
- ✓ Considere cómo controlará la distribución.
- ✓ Identifique los problemas de distribución y realice un plan de contingencia.

- Planificación de monitorización y del mantenimiento: Se debe registrar elementos y asegurarse de incluir en el informe final:
  - ✓ En cada modelo o descubrimiento, ¿qué factores o influencias necesita controlar?
  - ✓ ¿Cómo se puede medir y controlar la validez y precisión de cada modelo?
  - ✓ ¿Cómo se determina que un modelo ha “expirado”?
  - ✓ ¿Qué ocurre cuando un modelo expira? ¿Puede reconstruir el modelo con nuevos datos o tiene que realizar algunas modificaciones? ¿O por contra, las modificaciones son tantas que se requiere un nuevo proyecto de minería de datos?
  - ✓ ¿Se puede utilizar este modelo para problemas comerciales similares una vez expirado?
- Creación del informe final: Crear un informe final ayuda a comunicar los resultados. El informe debe incluir los siguientes elementos:
  - ✓ Una descripción detallada del problema original.
  - ✓ El procedimiento utilizado para realizar el proyecto de minería de datos.
  - ✓ El coste del proyecto.
  - ✓ Comentarios sobre las desviaciones del plan del proyecto original.
  - ✓ Un resumen de los resultados de minería de datos, incluyendo los modelos y los descubrimientos.

- ✓ Un resumen del plan propuesto para la distribución.

- Revisión del proyecto final: Es el paso final del método CRISP-DM y ofrece una oportunidad de formular las impresiones finales e incorporar los conocimientos adquiridos. Se debe realizar una breve entrevista con las personas implicadas en el proceso de minería de datos. Entre las cuestiones que debe tener en cuenta durante las entrevistas que realice se incluyen:
  - ✓ ¿Cuál es su impresión global del proyecto?
  - ✓ ¿Qué conocimientos ha adquirido durante el proceso de minería de datos en general y los datos disponibles?
  - ✓ ¿Qué partes del proyecto han funcionado correctamente? ¿Dónde han surgido las dificultades? ¿Existe algún tipo de información que le podría haber evitado confusiones?

2.2.5. Contexto tecnológico

2.2.5.1. Lenguaje de programación

Java

Java es un lenguaje de programación y una plataforma informática tanto para aplicaciones y sitios web, es un lenguaje de programación rápido y seguro y fiable. Dentro de su plataforma incluyen un ambiente de ejecución de programas denominado (JRE), el cual está formado por la máquina virtual de java (JVM), clases de núcleo de la plataforma Java y bibliotecas de la plataforma de soporte (Oracle, 2014).

R

R es un lenguaje de programación interpretado, de distribución libre, bajo Licencia GNU, y se mantiene en un ambiente para el cómputo estadístico y gráfico. Este software corre en distintas plataformas Linux, Windows, MacOS, e incluso en PlayStation 3. R no sólo es un sistema estadístico es también un ambiente en el que se aplican técnicas estadísticas (Santana & Farfán, 2014).

Python

Python es un lenguaje de programación interpretado, orientado a objetos de alto nivel con semántica dinámica. Contiene funciones de integradas para distintas estructuras de datos, combinado con asignación de tipos dinámicos la hacen atractiva para el desarrollo de aplicaciones rápidas. Python tiene una

sintaxis simple y fácil de aprender, soporta el uso de módulos y paquetes que permiten a los programas hacer rehúso de código (Python.org, 2010).

2.2.5.2. Entorno de desarrollo

Weka

El programa Weka es una colección de algoritmos de machine learning y herramientas de procesamiento de datos de última generación. Está diseñado para probar rápidamente métodos existentes en nuevos conjuntos de datos de maneras flexibles. Proporciona un amplio apoyo para todo el proceso de minería de datos experimentales, incluyendo la preparación de los datos de entrada, la evaluación de los esquemas de aprendizaje estadísticamente y la visualización de los datos de entrada y el resultado del aprendizaje. Además de una variedad de algoritmos de aprendizaje automático. Se accede por medio de una interfaz común para que el usuario pueda comparar diferentes métodos e identificar aquellos que son más apropiados para el problema planteado (Ian & Eibe, 2005, p. 365).

Weka fue desarrollado por la Universidad de Waikato en Nueva Zelanda y el nombre significa Waikato Environment for Knowledge Analysis, el sistema está escrito en Java y distribuido bajo términos GNU General Public Licence. Se puede ejecutar en cualquier plataforma y ha sido probada en sistemas operativos como Linux, Windows y Macintosh. Provee una interfaz uniforme para muchos algoritmos de machine learning acompañado con métodos de pre y post procesamiento y esquemas de evaluación de resultados de aprendizaje en un conjunto de datos dado. (Ian & Eibe, 2005, p. 366).

Netbeans

Netbeans es un ambiente integrado de desarrollo de código libre que permite el desarrollo de aplicaciones en sistemas operativos como Windows, Mac, Linux y Solaris. El IDE simplifica el desarrollo de aplicaciones web, empresariales, de escritorio y móviles que usan lenguaje Java y HTML5. También ofrece soporte para desarrollo de aplicaciones con lenguaje PHP y C/C++ (Oracle, 2016).

Apache Spark

Spark es una plataforma de computación de código abierto para análisis y procesos avanzados, que tiene muchas ventajas sobre Hadoop. Spark fue diseñado para soportar algoritmos iterativos que se pueden desarrollar sin escribir un conjunto de resultados cada vez que se procesa un dato. Apache

Spark tiene una velocidad de procesamiento cien veces más rápida que MapReduce.

Spark contiene un framework integrado para implementar análisis avanzados que incluye la librería MLlib, el motor gráfico GraphX, Spark Streaming y la herramienta de consulta Shark (Tirados, 2014).

2.2.5.3. Gestores de datos

SQL Server

SQL Server es un sistema de gestión de base de datos relacionales de Microsoft que está diseñado para el entorno empresarial. SQL Server se ejecuta en T-SQL – Transact SQL – que es un conjunto de extensiones de programación de Sybase y Microsoft que añaden varias características a SQL estándar, incluye control de transacciones, excepción y manejo de errores, procesamiento de filas, así también cómo variables declaradas. (Rouse M, 2015).

Mongo DB

MongoDB es una base de datos de propósito general potente, flexible y escalable. Combina capacidad de escalar con características tales como índices secundarios, consultas de rango, clasificación, agregaciones e índices geoespaciales.

MongoDB es una base de datos orientada a documentos no relacional. Una base de datos orientada a documentos reemplaza el concepto de fila con un modelo más flexible llamado documento. Permite realizar documentos anidados y arreglos, el enfoque orientado a documentos hace posible representar relaciones jerárquicas complejas con un simple registro. MongoDB está diseñado para escalar de manera muy rápida (Chodorow, 2013).

CAPÍTULO 3. HIPÓTESIS

3.1. Formulación de la Hipótesis

Considerando la formulación de la siguiente hipótesis:

El desarrollo de una aplicación basada en un modelo de machine learning contribuye a mejorar la evaluación de préstamos crediticios.

3.2. Operacionalización de variables

Tabla Nº 1 Operacionalización de la variable dependiente

Variable Dependiente	Definición conceptual	Definición operacional	Dimensión	Indicador
Evaluación de préstamos crediticios	Consiste en formarse un juicio acerca de si la personalidad, capacidad y avales de un postulante garantizan la buena utilización y el reembolso del préstamo.	Es una operación financiera donde un acreedor (entidad financiera) evalúa la solicitud de un préstamo para identificar el riesgo del crédito y la rentabilidad del mismo según una cantidad determinada de dinero concedida a una persona (deudor), en la cual éste se compromete a devolver la cantidad solicitada en un plazo definido.	Rentabilidad	Porcentaje de dinero ganado por préstamos crediticios clasificados
			Riesgo crediticio	Cantidad de dinero perdido por préstamos crediticios clasificados
			Tiempo	Tiempo promedio en días para aprobar un préstamo crediticio

Fuente: Elaboración propia

Tabla Nº 2 Operacionalización de la variable independiente

Variable Independiente	Definición conceptual	Definición operacional	Dimensión	Indicador
Aplicación informática basada en un modelo de machine learning	Una aplicación informática de machine learning es un conjunto de herramientas diseñadas para realizar tareas basadas en un modelo matemático de aprendizaje que permite que la aplicación pueda mostrar resultados aproximados sobre lo que se está analizando. El modelo puede ser predictivo, descriptivo o ambos.	La aplicación informática de machine learning permite al usuario ingresar los datos o perfil de un cliente que solicita un préstamo, y la aplicación a través de algoritmos evalúa si el cliente está apto para recibir dicho préstamo.	Sensibilidad	Porcentaje de préstamos crediticios clasificados como aprobados
			Especificidad	Porcentaje de préstamos crediticios clasificados como rechazados
			Eficacia	Porcentaje de préstamos crediticios clasificados de manera correcta

Fuente: Elaboración propia

CAPÍTULO 4. DESARROLLO

4.1. Comprensión del negocio

4.1.1. Determinar los objetivos del negocio

Los objetivos del negocio considerados son aquellos que están directamente asociados con los objetivos del desarrollo del modelo de machine learning.

4.1.1.1. Objetivos del negocio

Dentro de los objetivos identificados para el presente proyecto se encuentran los siguientes:

Tabla N° 3 Objetivos del Negocio

CÓDIGO	OBJETIVOS DEL NEGOCIO
OBNG-01	Brindar un servicio de calidad con tecnologías modernas para los procesos que involucran a los clientes.
OBNG-02	Mejorar el proceso de evaluación de préstamos en un 50% para minimizar el índice de morosidad de los clientes
OBNG-03	Mejorar el proceso de evaluación de préstamos para disminuir en un 50% el tiempo de aprobación de un préstamo crediticio

4.1.1.2. Criterios de éxito del negocio

Para cada objetivo se han definido los criterios de aceptación.

Tabla N° 4 Criterios de Aceptación

OBJETIVO	CRITERIO DE ACEPTACIÓN
OBNG-01	Implantar puesta en entorno de desarrollo para validar y verificar el funcionamiento de los procesos.
OBNG-02	Tener implementado un sistema informático automático que brinde un resultado de evaluación mayor al 70%
OBNG-03	Tener una plataforma o sistema válido, que brinde un nivel de eficiencia mayor al 80%

4.1.2. Evaluación de la situación

En este punto se describen los recursos, requisitos, supuestos, restricciones, riesgos, terminologías y costes del proyecto.

4.1.2.1. Inventario de recursos

El inventario contiene los recursos humanos y los orígenes de datos

Tabla N° 5 Recursos Humanos

CÓDIGO	RECURSOS HUMANOS
RH-01	Jorge Junior Rodríguez Castillo
RH-02	Milagros Madeleine Miñano Ochoa

Tabla N° 6 Fuentes de Datos

CÓDIGO	FUENTE DE DATOS
FD-01	Base de datos SQL Server 2012 con la información de los prestamos crediticios de los clientes

4.1.2.2. Requisitos, supuestos y restricciones

Tabla N° 7 Requisitos del Proyecto

CÓDIGO	REQUISITOS DEL PROYECTO
RQ-01	Contar con información de los préstamos crediticios de un periodo mínimo de 2 años.
RQ-02	Contar con información sobre el proceso de préstamos crediticios del negocio
RQ-03	Tener acceso de lectura de las fuentes de datos
RQ-04	Utilizar tecnología que permite obtener resultados confiables

Tabla N° 8 Supuestos del Proyecto

CÓDIGO	SUPUESTOS DEL PROYECTO
ST-01	Se cuenta con datos relevantes para el desarrollo del proyecto
ST-02	La aplicación informática sólo cumple con uno de los indicadores
ST-03	El costo del desarrollo de la aplicación informática es muy elevado para el negocio

ST-04	La entidad financiera no brinda información suficiente del proceso de préstamos crediticios.

Tabla N° 9 Restricciones del Proyecto

CÓDIGO	RESTRICCIONES DEL PROYECTO
RT-01	No contar con acceso a datos actualizados.

4.1.2.3. Riesgos y contingencias

El desarrollo del proyecto se puede ver afectado por algunos riesgos, para los cuales es necesario determinar una solución.

Tabla N° 10 Riesgos y Contingencia

CÓDIGO	RIESGOS	CONTINGENCIA
RG-01	La información con la que se cuenta no esté completa	Fijar un periodo de tiempo con la información que se cuenta para realizar un modelo más confiable
RG-02	Datos pocos relevantes.	

4.1.2.4. Terminología

Tabla N° 11 Términos del Proyecto

CÓDIGO	TÉRMINO
ML	Machine Learning
CRISP	Cross-Industry Standard Process for Data Mining
RL	Regresión Logística
features	Características

4.1.2.5. Costes

Tabla N° 12 Costos de Hardware

RECURSOS DE HARDWARE			
Equipo	**Cantidad**	**Precio**	**Total (depreciación 3 años)**
Laptop	2	2700,00	1800,00
Impresora/Scanner	1	300,00	100,00

TOTAL	**S/. 1900,00**

Tabla N° 13 Costos de Software

RECURSOS DE SOFTWARE			
Descripción	**Cantidad**	**Precio**	**Total**
Windows 10	1	215,82	215,82
Microsoft Office 2013	1	120,00	120,00
Weka	1	0,00	0,00
Rjava	1	0,00	0,00
TOTAL			**S/. 335,82**

Tabla N° 14 Costos de Internet

INTERNET			
Descripción	**Precio (mes)**	**Meses**	**Total**
Internet Claro de 4 mbps	88,00	8	704,00
TOTAL			**S/. 704,00**

Tabla N° 15 Costos de Recursos Humanos

RECURSOS HUMANOS			
Personal	**Meses**	**Precio**	**Total**
Analista/Programador	8	750,00	6000,00
Tester	8	750,00	6000,00
TOTAL			**S/. 3.600,00**

Tabla N° 16 Costos de Materiales

Recursos de Materiales				
Descripción	**Cantidad**	**Unidad**	**Precio**	**Total**
Lapiceros	1	Unidad	0,50	0,50
Folder Manila	5	Unidad	0,60	3,00
Tinta líquida Negro EPSON	1	Unidad	32,90	32,90
Tinta líquida Color EPSON	3	Unidad	32,90	98,70
Papel Bond A4	2	Millar	25,00	50,00
CD-RW	2	Unidad	1,00	2,00

Empastado	1	Unidad	20,00	20,00
TOTAL				**S/. 207,10**

Tabla N° 17 Ahorro en personal

Personal	Sueldo	Tiempo Ahorrado	Monto	Total
	Hora (S/.)	Mensual (horas)		
Analista de crédito	12,50	40,00	500,00	6000,00
TOTAL			500,00	**S/. 6000,00**

4.1.3. Determinar los objetivos

Los objetivos están divididos por objetivo general y específicos

4.1.3.1. Objetivo general

Determinar la mejora de la evaluación de préstamos crediticios mediante una aplicación informática basada en un modelo de machine learning.

4.1.3.2. Objetivos específicos

- Aumentar el porcentaje de dinero ganado por los préstamos crediticios clasificados.
- Disminuir la cantidad de dinero perdido por préstamos crediticios clasificados.
- Disminuir el tiempo promedio empleado en realizar el proceso de evaluación de préstamos.
- Aumentar el porcentaje de préstamos crediticios clasificados de manera correcta por medio del uso de la aplicación informática con el modelo de machine learning.

4.1.4. Realizar el plan del proyecto

4.1.4.1. Conformación del equipo

Tabla N° 18 Miembros del Equipo

Nº	MIEMBRO	ROL
01	Jorge Rodríguez Castillo	• Analista / Programador • Tester
02	Milagros Miñano	• Analista / Programador • Tester
03	Ing. Orlando Salazar	Asesor

4.1.4.2. Requerimientos

4.1.4.2.1. Requerimientos funcionales

- La aplicación informática debe permitir la evaluación de préstamos crediticios.

4.1.4.2.2. Requerimientos no funcionales

- Las interfaces de la aplicación informáticas deben ser amigables.
- La aplicación informática debe ser de alto rendimiento.
- La evaluación de los préstamos crediticios debe realizarse de forma rápida.
- La evaluación de los préstamos crediticios debe ser eficiente.

4.1.4.3. Fases de desarrollo

A continuación las fases de desarrollo:

Tabla N° 19 Comprensión del negocio

FASE 01
Nombre de la fase: Comprensión del negocio
Encargados: Milagros Miñano Ochoa Jorge Rodríguez Castillo
Descripción: Se establece los objetivos del negocio y la evaluación del negocio. Tareas de la fase de desarrollo: • Determinar objetivos del negocio • Evaluación de la situación • Determinar los objetivos • Realizar el plan de proyecto

Tabla N° 20 Comprensión de los datos

FASE 02
Nombre de la fase: Comprensión de los datos
Encargados: Milagros Miñano Ochoa Jorge Rodríguez Castillo

Descripción: Acceder y explorar los datos. Tareas de la fase de desarrollo: • Recopilación de datos iniciales • Descripción de los datos • Exploración de los datos • Verificar la calidad de los datos

Tabla N° 21 Preparación de los datos

FASE 03
Nombre de la fase: Preparación de los datos
Encargados: Milagros Miñano Ochoa Jorge Rodríguez Castillo
Descripción: Seleccionar, limpiar e integrar los datos. Tareas de la fase de desarrollo • Seleccionar los datos • Limpiar los datos • Construir los datos • Integrar los datos • Formateo de los datos

Tabla N° 22 Modelado

FASE 04
Nombre de la fase: Modelado
Encargados: Milagros Miñano Ochoa Jorge Rodríguez Castillo
Descripción: Determinar y aplicar un modelo. Tareas de la fase de desarrollo: • Seleccionar la técnica de modelado. • Generar un diseño de comprobación. • Arquitectura del modelo. • Generar los modelos.

Tabla N° 23 Evaluación del modelo

FASE 05
Nombre de la fase: Evaluación del modelo
Encargados: Milagros Miñano Ochoa Jorge Rodríguez Castillo
Descripción: Evaluar los resultados obtenidos, según lo planteado en la primera iteración.

Tabla N° 24 Implantación

FASE 06
Nombre de la fase: Implantación
Encargados: Milagros Miñano Ochoa Jorge Rodríguez Castillo
Descripción: Implantación en entorno de desarrollo y/o producción. Tareas de la fase de desarrollo: • Planear la implantación • Planear la monitorización y mantenimiento • Producir el informe final • Revisar el proyecto

4.1.4.4. Planificación inicial

Se define la prioridad (Bajo, Media o Alta según la importancia que tenga), riesgo (Bajo, Medio o Alto es la probabilidad de fallo en cada fase), esfuerzo (Se califica Bajo, Medio o Alto según el tiempo y trabajo que nos demandará en desarrollar la fase) e iteración (Es el orden de la implementación de cada fase, se califica del 1 al 6) de cada historia de usuario.

Tabla N° 25 Planificación inicial

Nº	Fase de desarrollo	Prioridad	Riesgo	Esfuerzo	Iteración
01	Comprensión del negocio	Alta	Bajo	Bajo	1

02	Comprensión de los datos	Alta	Medio	Alto	2
03	Preparación de los datos	Alta	Alto	Alto	3
04	Modelado	Alta	Alto	Alto	4
05	Evaluación del modelado	Alta	Medio	Medio	5
06	Implantación	Alta	Medio	Medio	6

4.1.4.5. Velocidad del proyecto

De acuerdo a las ponderaciones de la prioridad, riesgo y esfuerzo se ha estimado el tiempo de desarrollo de cada historia.

Tabla N° 26 Tiempo estimado en el desarrollo

Nº	Fase de desarrollo	Tiempo estimado
01	Comprensión del negocio	4 días
02	Comprensión de los datos	15 días
03	Preparación de los datos	20 días
04	Modelado	12 días
05	Evaluación del modelado	8 días
06	Implantación	4 días
Total		63 días

4.2. Comprensión de los datos

4.2.1. Recopilación de datos iniciales

La recopilación de datos se realizó a través de las siguientes fuentes:

- Base de datos – BDPRESTAMOSFNC: Esta base de datos contiene información sobre la solicitud de préstamos crediticios, los datos del cliente. En algunos casos, los registros no están completos. Se cuenta con registros desde el 2014 hasta el 2015.

4.2.2. Descripción de los datos

- Para el desarrollo del proyecto, se utilizan los datos dentro del periodo de enero hasta diciembre de los años 2014 y 2015, contando con un total de 1460 préstamos crediticios.
- Diccionario de datos (solo se considera los datos necesarios para el desarrollo del proyecto).

Tabla N° 27 Descripción de Tabla DocumentoGenerado

Nombre de la tabla: DocumentoGenerado				
Datos	**Descripción**	**Tipo de dato**	**Requerido**	**Restricción**
IdOficina	Identificador de la oficina	char(2)	Sí	Clave Primaria
IdDocumento	Identificador del documento	char(4)	Sí	Clave Primaria
NroDocumento	Número de documento	char(7)	Sí	Clave Primaria
TipoMoneda	Tipo de moneda	char(1)	Sí	Clave Primaria
IdEstadoDocumento	Identificador del estado del documento	char(2)	Sí	Clave foránea
IdOperacion	Identificador de la operación	char(5)	Sí	Clave foránea
FechaDocumento	Fecha de registro del documento	datetime	Sí	
HoraDocumento	Hora de registro del documento	char(8)	Sí	
IdPersona	Identificador del socio	char(7)	Sí	Clave foránea
GlosaFija	Glosa	varchar(100)	No	
FechaProceso	Fecha de proceso	datetime	Sí	
HoraProceso	Hora del proceso	char(7)	Sí	

Tabla N° 28 Descripción de Tabla DatosdeCredito

Nombre de la tabla: DatosdeCredito				
Datos	**Descripción**	**Tipo de dato**	**Requerido**	**Restricción**
IdOficina	Identificador de la oficina	char(2)	Sí	Clave Primaria
IdDocumento	Identificador del documento	char(4)	Sí	Clave Primaria
NroDocumento	Número del documento	char(7)	Sí	Clave Primaria
TipoMoneda	Tipo de moneda	char(1)	Sí	Clave Primaria
MontoSolicitado	Monto solicitado	money	Sí	
PlazoSolicitado	Plazo solicitado	int	Sí	
InteresSolicitado	Interés solicitado	money	Sí	
MontoAprobado	Monto aprobado	money	Sí	
PlazoAprobado	Plazo aprobado	int	Sí	
InteresAprobado	Interés aprobado	money	Sí	
IdAnalista	Identificador de analista	char(7)	Sí	
MontoAprobado	Monto aprobado	money	Sí	

Tabla N° 29 Descripción de Tabla Persona

Nombre de la tabla: Persona				
Datos	**Descripción**	**Tipo de dato**	**Requerido**	**Restricción**
IdPersona	Identificador de la persona	char(7)	Sí	Clave Primaria
FechaNacimiento	Fecha de nacimiento	datetime	Sí	
IngresoMensual	Monto de ingreso mensual de la persona	money	Sí	
Direccion	Dirección de la vivienda de la persona	varchar(650)	No	

Tabla N° 30 Descripción de Tabla PersonaNatural

Nombre de la tabla: PersonaNatural				
Datos	**Descripción**	**Tipo de dato**	**Requerido**	**Restricción**
IdPersona	Identificador de la persona	char(7)	Sí	Clave Primaria
PrimerApellido	Primer apellido	varchar(35)	Sí	
SegundoApellido	Segundo apellido	varchar(35)	Sí	
Nombres	Nombres	varchar(40)	Sí	
Sexo	Sexo de la persona	char(1)	Sí	
IdEstadoCivil	Identificador del estado civil	char(4)	Sí	

Tabla N° 31 Descripción de Tabla PersonaJurídica

Nombre de la tabla: PersonaJurídica				
Datos	**Descripción**	**Tipo de dato**	**Requerido**	**Restricción**
IdPersona	Identificador de la persona	char(7)	Sí	Clave Primaria
IdUsuario	Identificador de usuario	char(6)	Sí	
NombreCorto	Nombre de la persona jurídica	varchar(15)	Sí	
RazonSocial	Razón social	varchar(80)	Sí	
FechaProceso	Fecha de proceso	datetime	Sí	
HoraProceso	Hora de proceso	char(7)	Sí	

DIAGRAMA DE BASE DE DATOS PARA EXTRACCIÓN DE INFORMACIÓN

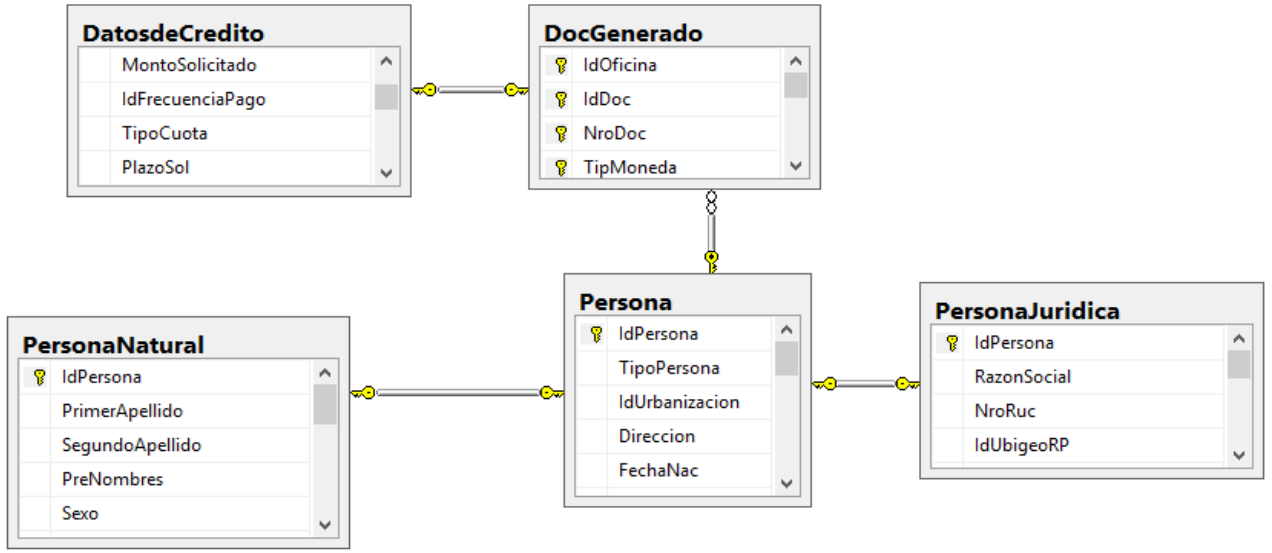


Figura N° 14 Diagrama de base de datos

Fuente: Elaboración propia

4.2.3. Verificar la calidad de los datos

En la siguiente lista se menciona los puntos considerados para verificar la calidad de los datos:

Tabla N° 32 Calidad de los datos

CÓDIGO	DESCRIPCIÓN
CLD-01	Los campos de las tablas no deben tener valores nulos.
CLD-02	Los campos de las tablas que guarda el monto solicitado o aprobado, no deben tener valor cero.
CLD-03	La información referente a la solicitud de datos no debe estar duplicada.

4.3. Preparación de los datos

4.3.1. Seleccionar los datos

De acuerdo con los datos que se muestran en el diccionario de datos (Tablas N° 27, 28, 29, 30, 31), se realiza un archivo con consultas en lenguaje Transact-SQL para seleccionar las columnas – características – y las filas – conjunto de datos – requeridas para ser usadas en el algoritmo de machine learning.

```
SELECT DG.*, DC.*, PN.IdPersona, PN.IdUsuario, CONCAT(PN.PreNombres,' ', PN.PrimerApellido, ' ', PN.SegundoApellido) as Nombres, PN.Sexo, PN.IdEstadoCivil, '' as RazonSocial, PN.FechaProceso, PN.HoraProceso, P.*, T.Descripcion as EstadoCivil, T2.Descripcion as Profesion
FROM DocGenerado DG INNER JOIN
     DatosdeCredito DC ON(DG.IdOficina = DC.IdOficina AND DG.IdDoc = DC.IdDoc AND DG.NroDoc = DC.NroDoc AND DG.TipMoneda = DC.TipMoneda) INNER JOIN      PersonaNatural PN ON(DC.IdPersonaSocio = PN.IdPersona) INNER JOIN
        Persona P ON (PN.IdPersona = P.IdPersona) LEFT JOIN
        Tablas T ON (T.IdTabla = PN.IdEstadoCivil) LEFT JOIN
        Tablas T2 ON (T2.IdTabla = PN.IdProfesion)
WHERE DG.IdOficina = '01'
  AND DG.IdDoc = '0007'
  AND DG.IdOpe = '00172'
  AND DG.TipMoneda = '1'
  AND DC.MontoSolicitado > 0
  AND YEAR(DG.FechaProceso) IN(2014,2015)
  AND DG.IdEstDoc NOT IN ('01', '05', '09', '10')
union
SELECT DG.*, DC.*, PJ.IdPersona, PJ.IdUsuario, PJ.NombreCorto as Nombres, NULL as Sexo, NULL as IdEstadoCivil, PJ.RazonSocial, PJ.FechaProceso, PJ.HoraProceso, P.*, NULL as EstadoCivil, NULL as Profesion
```

```
FROM DocGenerado DG INNER JOIN
     DatosdeCredito DC ON(DG.IdOficina = DC.IdOficina AND DG.IdDoc =
DC.IdDoc AND DG.NroDoc = DC.NroDoc AND DG.TipMoneda = DC.TipMoneda) INNER
JOIN
        PersonaJuridica PJ ON(DC.IdPersonaSocio = PJ.IdPersona) INNER
JOIN
        Persona P ON (PJ.IdPersona = P.IdPersona)
WHERE DG.IdOficina = '01'
  AND DG.IdDoc = '0007'
  AND DG.IdOpe = '00172'
  AND DG.TipMoneda = '1'
  AND DC.MontoSolicitado > 0
  AND YEAR(DG.FechaProceso) IN(2014,2015)
  AND DG.IdEstDoc NOT IN ('01', '05', '09', '10')
```

Figura N° 15 Selección de créditos aceptados
Fuente: Elaboración propia

Oficina	CodDoc	NumeroDocumento	Moneda	Operacion	Fecha	Hora	IdPersona	Sexo	IdEstadoCivil	RazonSocial	FechaProceso	HoraProceso
01	0007	0069014	1	00172	2013-12-10 00:00:00.000	08:51AM	0035838	M	0401		2012-06-28 00:00:00.000	10:37AM
01	0007	0069416	1	00172	2014-01-02 00:00:00.000	12:44PM	0039530	F	0401		2013-12-27 00:00:00.000	10:35AM
01	0007	0069420	1	00172	2014-01-02 00:00:00.000	03:56PM	0012517	F	0401		2006-10-24 00:00:00.000	06:13PM
01	0007	0069435	1	00172	2014-01-03 00:00:00.000	11:07AM	0002877	M	0401		2012-01-06 00:00:00.000	12:39PM
01	0007	0069457	1	00172	2014-01-04 00:00:00.000	10:08AM	0021249	F	0401		2008-11-17 00:00:00.000	05:45PM
01	0007	0069468	1	00172	2014-01-06 00:00:00.000	10:01AM	0042538	M	0401		2014-01-03 00:00:00.000	06:12PM
01	0007	0069523	1	00172	2014-01-08 00:00:00.000	09:40AM	0032338	M	0401		2015-05-15 00:00:00.000	11:02AM
01	0007	0069500	1	00172	2014-01-07 00:00:00.000	09:16AM	0010607	M	0401		2006-05-05 00:00:00.000	08:57AM
01	0007	0069554	1	00172	2014-01-09 00:00:00.000	03:37PM	0042609	M	0401		2014-01-08 00:00:00.000	04:39PM
01	0007	0069556	1	00172	2014-01-09 00:00:00.000	05:10PM	0042521	M	0401		2014-01-03 00:00:00.000	12:45PM
01	0007	0069564	1	00172	2014-01-10 00:00:00.000	10:05AM	0005359	F	0401		2009-12-28 00:00:00.000	05:46PM
01	0007	0069486	1	00172	2014-01-06 00:00:00.000	12:59PM	0005012	F	0401		2011-12-05 00:00:00.000	02:56PM
01	0007	0069488	1	00172	2014-01-06 00:00:00.000	01:25PM	0036288	M	0401		2012-08-10 00:00:00.000	10:35AM
01	0007	0069600	1	00172	2014-01-13 00:00:00.000	09:01AM	0020181	F	0401		2008-08-18 00:00:00.000	04:32PM
01	0007	0069601	1	00172	2014-01-13 00:00:00.000	09:05AM	0042625	M	0401		2014-01-10 00:00:00.000	11:46AM

Figura Nº 16 Información de la base de datos de los créditos aceptados
Fuente: Elaboración propia

```
SELECT DG.*, DC.*, PN.IdPersona, PN.IdUsuario, CONCAT(PN.PreNombres,' ',
PN.PrimerApellido, ' ', PN.SegundoApellido) as Nombres, PN.Sexo,
PN.IdEstadoCivil, '' as RazonSocial, PN.FechaProceso, PN.HoraProceso,
P.*, T.Descripcion as EstadoCivil, T2.Descripcion as Profesion
FROM DocGenerado DG INNER JOIN
     DatosdeCredito DC ON(DG.IdOficina = DC.IdOficina AND DG.IdDoc =
DC.IdDoc AND DG.NroDoc = DC.NroDoc AND DG.TipMoneda = DC.TipMoneda) INNER
JOIN
        PersonaNatural PN ON(DC.IdPersonaSocio = PN.IdPersona) INNER
JOIN
        Persona P ON (PN.IdPersona = P.IdPersona) LEFT JOIN
        Tablas T ON (T.IdTabla = PN.IdEstadoCivil) LEFT JOIN
        Tablas T2 ON (T2.IdTabla = PN.IdProfesion)
WHERE DG.IdOficina = '01'
  AND DG.IdDoc = '0007'
  AND DG.IdOpe = '00172'
  AND DG.TipMoneda = '1'
```

```
  AND DG.IdEstDoc = '10'
  AND DC.MontoSolicitado > 0
  AND YEAR(DG.FechaProceso) IN(2014,2015)
union
SELECT DG.*, DC.*, PJ.IdPersona, PJ.IdUsuario, PJ.NombreCorto as
Nombres, NULL as Sexo, NULL as IdEstadoCivil, PJ.RazonSocial, -- CREDITOS
RECHAZADOS
PJ.FechaProceso, PJ.HoraProceso, P.*, NULL as EstadoCivil, NULL as
Profesion
FROM DocGenerado DG INNER JOIN
     DatosdeCredito DC ON(DG.IdOficina = DC.IdOficina AND DG.IdDoc =
DC.IdDoc AND DG.NroDoc = DC.NroDoc AND DG.TipMoneda = DC.TipMoneda) INNER
JOIN
       PersonaJuridica PJ ON(DC.IdPersonaSocio = PJ.IdPersona) INNER
JOIN
       Persona P ON (PJ.IdPersona = P.IdPersona)
WHERE DG.IdOficina = '01'
  AND DG.IdDoc = '0007'
  AND DG.IdOpe = '00172'
  AND DG.TipMoneda = '1'
  AND DG.IdEstDoc = '10'
  AND DC.MontoSolicitado > 0
  AND YEAR(DG.FechaProceso) IN(2014,2015)
  order by P.IdPersona ASC
```

Figura N° 17 Selección de créditos rechazados
Fuente: Elaboración propia

Oficina	CodDoc	NumeroDocumento	Moneda	Operacion	Fecha	Hora	IdPersona	Sexo	IdEstadoCivil	RazonSocial	FechaProceso	HoraProceso	IdPersona	TipoPersona
01	0007	0078199	1	00172	2015-05-27 00:00:00.000	03:13PM	0000120	F	0402		2011-05-16 00:00:00.000	05:33PM	0000120	N
01	0007	0070004	1	00172	2014-02-04 00:00:00.000	12:11PM	0000382	F	0405		2011-01-24 00:00:00.000	12:41PM	0000382	N
01	0007	0070956	1	00172	2014-03-25 00:00:00.000	11:16AM	0001575	F	0402		2009-10-14 00:00:00.000	05:42PM	0001575	N
01	0007	0072521	1	00172	2014-06-27 00:00:00.000	04:37PM	0002622	F	0403		2007-02-05 00:00:00.000	09:17AM	0002622	N
01	0007	0072511	1	00172	2014-06-27 00:00:00.000	11:20AM	0003867	M	0401		2015-09-09 00:00:00.000	11:54AM	0003867	N
01	0007	0075534	1	00172	2014-12-24 00:00:00.000	11:09AM	0005296	F	0403		2015-09-04 00:00:00.000	12:24PM	0005296	N
01	0007	0069691	1	00172	2014-01-17 00:00:00.000	11:56AM	0007162	F	0401		2006-08-25 00:00:00.000	04:44PM	0007162	N
01	0007	0078353	1	00172	2015-06-05 00:00:00.000	11:32AM	0007455	M	0401		2009-10-22 00:00:00.000	04:35PM	0007455	N
01	0007	0078708	1	00172	2015-06-26 00:00:00.000	10:30AM	0009631	F	0401		2015-08-11 00:00:00.000	04:31PM	0009631	N
01	0007	0076960	1	00172	2015-03-13 00:00:00.000	02:52PM	0017005	F	0401		2007-12-21 00:00:00.000	05:09PM	0017005	N
01	0007	0076690	1	00172	2015-02-28 00:00:00.000	10:16AM	0017271	M	0402		2013-10-24 00:00:00.000	02:06PM	0017271	N
01	0007	0070689	1	00172	2014-03-10 00:00:00.000	11:50AM	0020181	F	0401		2008-08-18 00:00:00.000	04:32PM	0020181	N
01	0007	0077925	1	00172	2015-05-11 00:00:00.000	10:28AM	0023814	F	0401		2015-05-12 00:00:00.000	09:50AM	0023814	N
01	0007	0078577	1	00172	2015-06-18 00:00:00.000	04:08PM	0024906	M	0401		2009-09-18 00:00:00.000	04:37PM	0024906	N

Figura Nº 18 Información de la base de datos de los créditos rechazados
Fuente: Elaboración propia

El resultado obtenido es una nueva fuente de datos (BD_PRESTAMOS) que contiene únicamente las filas y columnas seleccionadas. A continuación los campos seleccionados:

Tabla N° 33 Descripción de la Tabla Prestamos

Nombre de la tabla: Prestamos				
Datos	**Descripción**	**Tipo de dato**	**Requerido**	**Restricción**
idPrestamo	Identificador de la persona	int	Sí	Clave Primaria
idOficina	Identificador de oficina	char(2)	Sí	
idDocumento	Identificador de documento	char(4)	Sí	
nroDocumento	Número de documento	nchar(7)	Sí	
tipoMoneda	Tipo de moneda	char(1)	Sí	
codigoOperacion	Código de la operación	char(5)	Sí	
fechaDocumento	Fecha de documento	datetime	Sí	
horaDocumento	Hora de registro del documento	char(7)	Sí	
fechaCambioDocumento	Fecha de aprobación del documento	datetime	Sí	
idPersona	Identificador de la persona	char(7)	Sí	
glosaFija	Glosa	varchar(300)	Sí	
idUsuario	Identificador de usuario	char(6)	Sí	
fechaProceso	Fecha de proceso	datetime	Sí	
horaProceso	Hora del proceso	char(7)	Sí	
montoSolicitado	Monto solicitado en el préstamo	money	Sí	
plazoSolicitado	Plazo solicitado	smallint	Sí	
interesSolicitado	Interés solicitado	money	Sí	
montoAprobado	Monto aprobado	money	Sí	

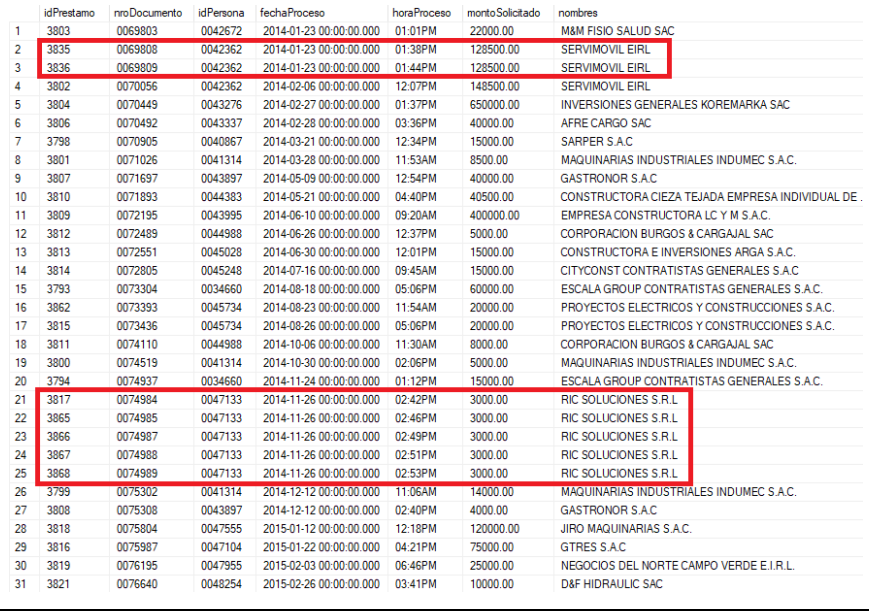
plazoAprobado	Plazo aprobado	smallint	Sí	
interesAprobado	Interés aprobado	money	Sí	
idAnalista	Identificador del analista	char(7)	Sí	
fechaDesembolso	Fecha de desembolso	datetime	Sí	
nombres	Nombres	varchar(300)	Sí	
sexo	Sexo	char(1)	Sí	
estadoCivil	Estado civil	char(18)	Sí	
tipoPersona	Tipo de persona	char(1)	Sí	
dirección	Dirección	varchar(300)	Sí	
fechaNacimiento	Fecha de nacimiento	datetime	Sí	
tipoVivienda	Tipo de vivienda	char(30)	Sí	
ingresoMensual	Ingreso Mensual	money	Sí	
Profesión	Profesión	varchar(80)	Sí	
estadoDocumento	Estado del documento	char(2)	Sí	
aceptado	Indica si el préstamo solicitado ha sido aceptado o rechazado.	smallint	Sí	

4.3.2. Limpiar los datos

La limpieza de datos se ha realizado a través de tres técnicas:

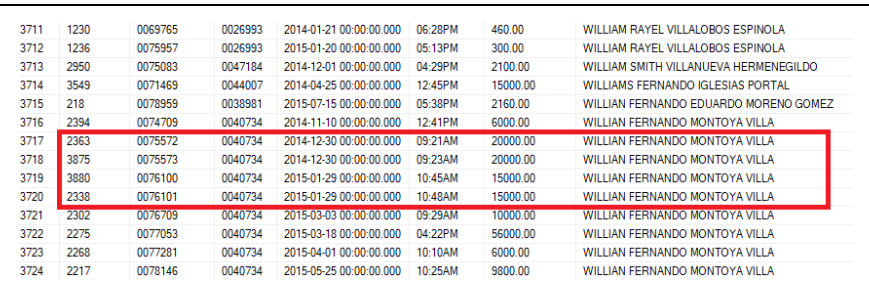
4.3.2.1. Verificación de duplicidad

Este proceso se ha realizado manualmente, para identificar la duplicidad en los registros se ha verificado si el nombre, monto solicitado, fecha y hora de proceso son iguales. En la base datos BD_PRESTAMOS se encontró con este tipo de inconsistencia y se dejó el último registro, como se puede observar en las imágenes.



	idPrestamo	nroDocumento	idPersona	fechaProceso	horaProceso	montoSolicitado	nombres
1	3803	0069803	0042672	2014-01-23 00:00:00.000	01:01PM	22000.00	M&M FISIO SALUD SAC
2	3835	0069808	0042362	2014-01-23 00:00:00.000	01:38PM	128500.00	SERVIMOVIL EIRL
3	3836	0069809	0042362	2014-01-23 00:00:00.000	01:44PM	128500.00	SERVIMOVIL EIRL
4	3802	0070056	0042362	2014-02-06 00:00:00.000	12:07PM	148500.00	SERVIMOVIL EIRL
5	3804	0070449	0043276	2014-02-27 00:00:00.000	01:37PM	650000.00	INVERSIONES GENERALES KOREMARKA SAC
6	3806	0070492	0043337	2014-02-28 00:00:00.000	03:36PM	40000.00	AFRE CARGO SAC
7	3798	0070905	0040867	2014-03-21 00:00:00.000	12:34PM	15000.00	SARPER S.A.C
8	3801	0071026	0041314	2014-03-28 00:00:00.000	11:53AM	8500.00	MAQUINARIAS INDUSTRIALES INDUMEC S.A.C.
9	3807	0071697	0043897	2014-05-09 00:00:00.000	12:54PM	40000.00	GASTRONOR S.A.C
10	3810	0071893	0044383	2014-05-21 00:00:00.000	04:40PM	40500.00	CONSTRUCTORA CIEZA TEJADA EMPRESA INDIVIDUAL DE .
11	3809	0072195	0043995	2014-06-10 00:00:00.000	09:20AM	400000.00	EMPRESA CONSTRUCTORA LC Y M S.A.C.
12	3812	0072489	0044988	2014-06-26 00:00:00.000	12:37PM	5000.00	CORPORACION BURGOS & CARGAJAL SAC
13	3813	0072551	0045028	2014-06-30 00:00:00.000	12:01PM	15000.00	CONSTRUCTORA E INVERSIONES ARGA S.A.C.
14	3814	0072805	0045248	2014-07-16 00:00:00.000	09:45AM	15000.00	CITYCONST CONTRATISTAS GENERALES S.A.C
15	3793	0073304	0034660	2014-08-18 00:00:00.000	05:06PM	60000.00	ESCALA GROUP CONTRATISTAS GENERALES S.A.C.
16	3862	0073393	0045734	2014-08-23 00:00:00.000	11:54AM	20000.00	PROYECTOS ELECTRICOS Y CONSTRUCCIONES S.A.C.
17	3815	0073436	0045734	2014-08-26 00:00:00.000	05:06PM	20000.00	PROYECTOS ELECTRICOS Y CONSTRUCCIONES S.A.C.
18	3811	0074110	0044988	2014-10-06 00:00:00.000	11:30AM	8000.00	CORPORACION BURGOS & CARGAJAL SAC
19	3800	0074519	0041314	2014-10-30 00:00:00.000	02:06PM	5000.00	MAQUINARIAS INDUSTRIALES INDUMEC S.A.C.
20	3794	0074937	0034660	2014-11-24 00:00:00.000	01:12PM	15000.00	ESCALA GROUP CONTRATISTAS GENERALES S.A.C.
21	3817	0074984	0047133	2014-11-26 00:00:00.000	02:42PM	3000.00	RIC SOLUCIONES S.R.L
22	3865	0074985	0047133	2014-11-26 00:00:00.000	02:46PM	3000.00	RIC SOLUCIONES S.R.L
23	3866	0074987	0047133	2014-11-26 00:00:00.000	02:49PM	3000.00	RIC SOLUCIONES S.R.L
24	3867	0074988	0047133	2014-11-26 00:00:00.000	02:51PM	3000.00	RIC SOLUCIONES S.R.L
25	3868	0074989	0047133	2014-11-26 00:00:00.000	02:53PM	3000.00	RIC SOLUCIONES S.R.L
26	3799	0075302	0041314	2014-12-12 00:00:00.000	11:06AM	14000.00	MAQUINARIAS INDUSTRIALES INDUMEC S.A.C.
27	3808	0075308	0043897	2014-12-12 00:00:00.000	02:40PM	4000.00	GASTRONOR S.A.C
28	3818	0075804	0047555	2015-01-12 00:00:00.000	12:18PM	120000.00	JIRO MAQUINARIAS S.A.C.
29	3816	0075987	0047104	2015-01-22 00:00:00.000	04:21PM	75000.00	GTRES S.A.C
30	3819	0076195	0047955	2015-02-03 00:00:00.000	06:46PM	25000.00	NEGOCIOS DEL NORTE CAMPO VERDE E.I.R.L.
31	3821	0076640	0048254	2015-02-26 00:00:00.000	03:41PM	10000.00	D&F HIDRAULIC SAC

Figura Nº 19 Duplicidad de créditos aceptados
Fuente: Elaboración propia



3711	1230	0069765	0026993	2014-01-21 00:00:00.000	06:28PM	460.00	WILLIAM RAYEL VILLALOBOS ESPINOLA
3712	1236	0075957	0026993	2015-01-20 00:00:00.000	05:13PM	300.00	WILLIAM RAYEL VILLALOBOS ESPINOLA
3713	2950	0075083	0047184	2014-12-01 00:00:00.000	04:29PM	2100.00	WILLIAM SMITH VILLANUEVA HERMENEGILDO
3714	3549	0071469	0044007	2014-04-25 00:00:00.000	12:45PM	15000.00	WILLIAMS FERNANDO IGLESIAS PORTAL
3715	218	0078959	0038981	2015-07-15 00:00:00.000	05:38PM	2160.00	WILLIAN FERNANDO EDUARDO MORENO GOMEZ
3716	2394	0074709	0040734	2014-11-10 00:00:00.000	12:41PM	6000.00	WILLIAN FERNANDO MONTOYA VILLA
3717	2363	0075572	0040734	2014-12-30 00:00:00.000	09:21AM	20000.00	WILLIAN FERNANDO MONTOYA VILLA
3718	3875	0075573	0040734	2014-12-30 00:00:00.000	09:23AM	20000.00	WILLIAN FERNANDO MONTOYA VILLA
3719	3880	0076100	0040734	2015-01-29 00:00:00.000	10:45AM	15000.00	WILLIAN FERNANDO MONTOYA VILLA
3720	2338	0076101	0040734	2015-01-29 00:00:00.000	10:48AM	15000.00	WILLIAN FERNANDO MONTOYA VILLA
3721	2302	0076709	0040734	2015-03-03 00:00:00.000	09:29AM	10000.00	WILLIAN FERNANDO MONTOYA VILLA
3722	2275	0077053	0040734	2015-03-18 00:00:00.000	04:22PM	56000.00	WILLIAN FERNANDO MONTOYA VILLA
3723	2268	0077281	0040734	2015-04-01 00:00:00.000	10:10AM	6000.00	WILLIAN FERNANDO MONTOYA VILLA
3724	2217	0078146	0040734	2015-05-25 00:00:00.000	10:25AM	9800.00	WILLIAN FERNANDO MONTOYA VILLA

Figura Nº 20 Duplicidad de créditos rechazados
Fuente: Elaboración propia

Para eliminar los registros duplicados se utilizó la siguiente consulta:

```
DELETE FROM Prestamos
WHERE idPersona in(
SELECT distinct idPersona
FROM Prestamos
GROUP BY idPersona, fechaProceso, montoSolicitado
HAVING count(*) > 1)
```

Figura N° 21 Consulta para eliminar registro duplicados
Fuente: Elaboración propia

4.3.2.2. Exclusión de características

Al analizar la base datos BD_PRESTAMOS se determinó que existen características que no son necesarias para el desarrollo del proyecto, por tal motivo, se realizó la exclusión de características. Las características que se utilizará finalmente son las siguientes:

Tabla N° 34 Exclusión de características

Datos	Descripción
idOficina	Identificador de oficina
idDocumento	Identificador de documento
nroDocumento	Número de documento
tipoMoneda	Tipo de moneda
codigoOperacion	Código de la operación
idPersona	Identificador de la persona
glosaFija	Glosa
idUsuario	Identificador de usuario
plazoSolicitado	Plazo solicitado
montoAprobado	Monto aprobado
plazoAprobado	Plazo aprobado
interesAprobado	Interés aprobado
idAnalista	Identificador del analista
fechaDesembolso	Fecha de desembolso
nombres	Nombres
tipoPersona	Tipo de persona
dirección	Dirección
fechaNacimiento	Fecha de nacimiento
tipoVivienda	Tipo de vivienda
profesion	Profesión
estadoDocumento	Estado del documento

A través del siguiente script se creó un nuevo conjunto de datos, que serán utilizados en el modelado.

```
INSERT INTO PrestamosModelo
(   idPrestamoModelo
  , montoSolicitado
  , interesSolicitado
  , ingresoMensual
  , sexo
  , estadoCivil
  , fechaNacimiento
  , fechaProceso
  , aceptado
)
SELECT idPrestamo
        , montoSolicitado
        , interesSolicitado
        , ingresoMensual
        , sexo
        , estadoCivil
        , fechaNacimiento
        , fechaProceso
        , aceptado

FROM Prestamos
```

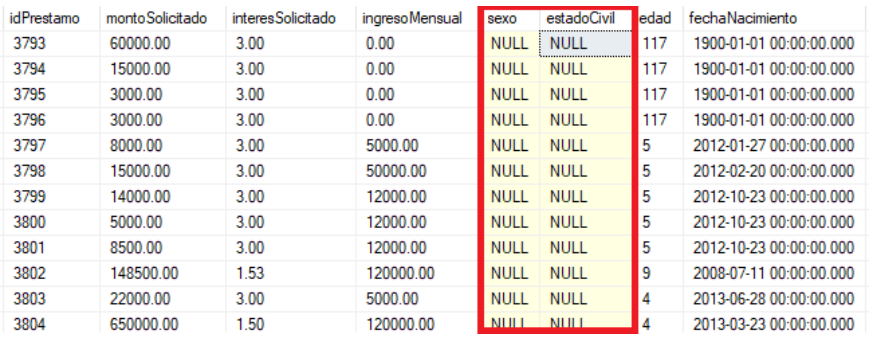
Figura N° 22 Consulta para creación de nuevo conjunto de datos
Fuente: Elaboración propia

4.3.2.3. Asignación de valores

En las solicitudes de préstamos de personas jurídicas se puede apreciar valores NULL, en las columnas sexo y estadoCivil. Por eso, en el nuevo conjunto de datos se le asigna a ambas columnas el valor N, el cual significa No Indica. A través de la siguiente consulta:

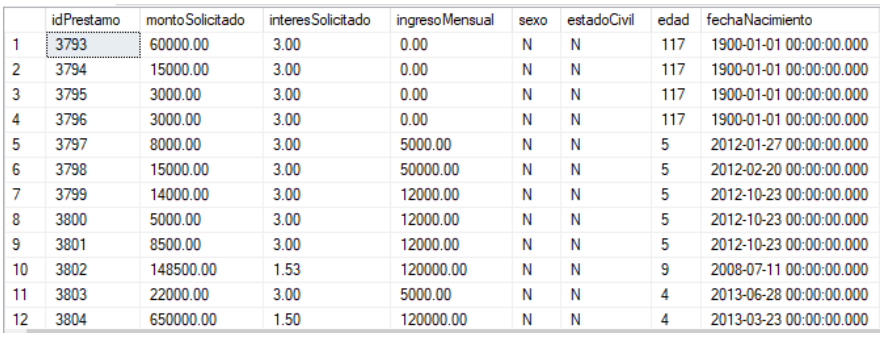
```
UPDATE P
SET P.sexo = 'N', P.estadoCivil = 'N'
FROM Prestamos_temp P
WHERE P.tipoPersona = 'J'
```

Figura N° 23 Consulta para asignación de nuevos valores
Fuente: Elaboración propia



idPrestamo	montoSolicitado	interesSolicitado	ingresoMensual	sexo	estadoCivil	edad	fechaNacimiento
3793	60000.00	3.00	0.00	NULL	NULL	117	1900-01-01 00:00:00.000
3794	15000.00	3.00	0.00	NULL	NULL	117	1900-01-01 00:00:00.000
3795	3000.00	3.00	0.00	NULL	NULL	117	1900-01-01 00:00:00.000
3796	3000.00	3.00	0.00	NULL	NULL	117	1900-01-01 00:00:00.000
3797	8000.00	3.00	5000.00	NULL	NULL	5	2012-01-27 00:00:00.000
3798	15000.00	3.00	50000.00	NULL	NULL	5	2012-02-20 00:00:00.000
3799	14000.00	3.00	12000.00	NULL	NULL	5	2012-10-23 00:00:00.000
3800	5000.00	3.00	12000.00	NULL	NULL	5	2012-10-23 00:00:00.000
3801	8500.00	3.00	12000.00	NULL	NULL	5	2012-10-23 00:00:00.000
3802	148500.00	1.53	120000.00	NULL	NULL	9	2008-07-11 00:00:00.000
3803	22000.00	3.00	5000.00	NULL	NULL	4	2013-06-28 00:00:00.000
3804	650000.00	1.50	120000.00	NULL	NULL	4	2013-03-23 00:00:00.000

Figura Nº 24 Registros con valores nulos
Fuente: Elaboración propia



	idPrestamo	montoSolicitado	interesSolicitado	ingresoMensual	sexo	estadoCivil	edad	fechaNacimiento
1	3793	60000.00	3.00	0.00	N	N	117	1900-01-01 00:00:00.000
2	3794	15000.00	3.00	0.00	N	N	117	1900-01-01 00:00:00.000
3	3795	3000.00	3.00	0.00	N	N	117	1900-01-01 00:00:00.000
4	3796	3000.00	3.00	0.00	N	N	117	1900-01-01 00:00:00.000
5	3797	8000.00	3.00	5000.00	N	N	5	2012-01-27 00:00:00.000
6	3798	15000.00	3.00	50000.00	N	N	5	2012-02-20 00:00:00.000
7	3799	14000.00	3.00	12000.00	N	N	5	2012-10-23 00:00:00.000
8	3800	5000.00	3.00	12000.00	N	N	5	2012-10-23 00:00:00.000
9	3801	8500.00	3.00	12000.00	N	N	5	2012-10-23 00:00:00.000
10	3802	148500.00	1.53	120000.00	N	N	9	2008-07-11 00:00:00.000
11	3803	22000.00	3.00	5000.00	N	N	4	2013-06-28 00:00:00.000
12	3804	650000.00	1.50	120000.00	N	N	4	2013-03-23 00:00:00.000

Figura Nº 25 Registros con valores asignados
Fuente: Elaboración propia

4.3.3. Construir los datos

El método utilizado es derivación de datos, pues es necesario contar con la edad de los solicitantes, por eso se construye una nueva columna en la base datos, Edad, tomando como referencia la fecha de nacimiento y la fecha en el que se empieza el proceso de solicitud de préstamo.

A través del siguiente procedimiento:

```
CREATE PROCEDURE spCrearCaracteristicaEdad
AS
BEGIN
   UPDATE PM
   SET PM.edad = DATEDIFF(YEAR, PM.fechaNacimiento, PM.fechaProceso)
   FROM PrestamosModelo PM
END
```

Figura N° 26 Procedimiento almacenado para calcular edad
Fuente: Elaboración propia

4.3.4. Integrar los datos

Se generó una nueva base de datos BD_PRESTAMOS, con dos tablas. La primera tabla contiene todos los datos que se seleccionaron de la base de datos original. La segunda tabla contiene los datos que serán utilizados para el modelado, estos datos se obtuvieron después de culminar con la limpieza de datos.

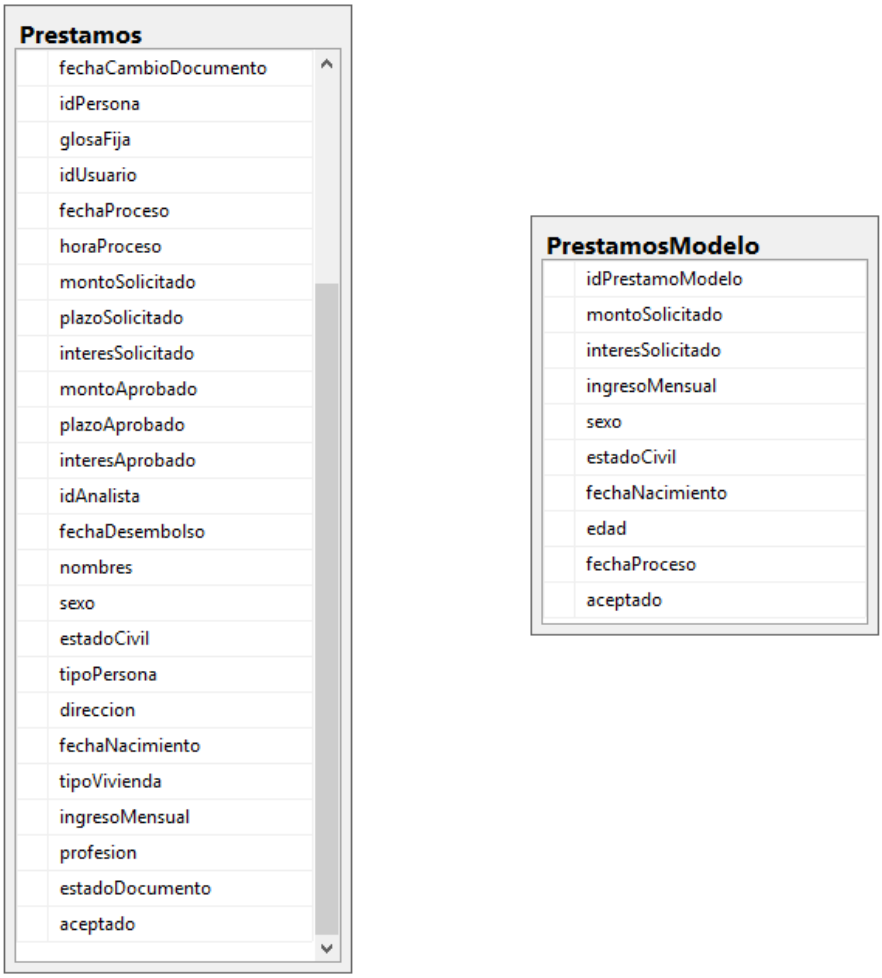


Figura N° 27 Tablas de la base de datos BD_PRESTAMOS
Fuente: Elaboración propia

4.3.5. Formato de los datos

El campo con la información referente a sexo y estado civil, ha sido codificado con valores numéricos ya que permite mayor rapidez en el procesamiento de datos. A través del siguiente procedimiento se realiza la asignación del nuevo valor.

```
CREATE PROCEDURE [dbo].[spActualizarColumnas]
AS
BEGIN
UPDATE P
```

```
SET P.sexo =  (CASE WHEN sexo = 'F' THEN 1 WHEN sexo = 'M' THEN 2 WHEN
sexo = 'N' THEN 0 END), P.estadoCivil = (   CASE WHEN estadoCivil =
RTRIM('N') OR estadoCivil = RTRIM('NO INDICA') THEN 0
              WHEN estadoCivil = RTRIM('CASADO (A)') THEN 1
              WHEN estadoCivil = RTRIM('CONVIVIENTE') THEN 2
              WHEN estadoCivil = RTRIM('DIVORCIADO (A)') THEN 3
              WHEN estadoCivil = RTRIM('SEPARADO') THEN 4
              WHEN estadoCivil = RTRIM('SOLTERO (A)') THEN 5
              WHEN estadoCivil = RTRIM('VIUDO(A)') THEN 6 END )
FROM PrestamosModelo P

END
```

Figura N° 28 Procedimiento almacenado para dar formato a los registros
Fuente: Elaboración propia

Tabla N° 35 Descripción de reemplazo de columna sexo

Columna sexo	
Valores	**Nuevos valores**
N	0
F	1
M	2

Tabla N° 36 Descripción de reemplazo de columna estadoCivil

Columna estadoCivil	
Valores	**Nuevos valores**
N	0
CASADO	1
CONVIVIENTE	2
DIVORCIADO	3
SEPARADO	4
SOLTERO	5
VIUDO	6

4.4. Modelado

4.4.1. Técnica del modelo

En el presente proyecto se ha considera el uso del algoritmo de machine learning denominado Regresión Logística. Este algoritmo se ajusta a los objetivos planteados en la investigación, debido a que es un algoritmo eficaz de clasificación, es rápido y sencillo de implementar.

4.4.2. Plan de pruebas del modelo

En el plan de pruebas se ha particionado los datos en dos conjuntos, uno de entrenamiento y otro de prueba de predicción, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado. Este plan es utilizado en el punto Evaluación del modelado.

Tabla N° 37 Partición de datos para plan de pruebas

	PORCENTAJE
DATA DE ENTRENAMIENTO	75%
DATA DE PRUEBAS	25%

Para las pruebas se ha utilizado la herramienta Weka, la cual requiere un formato específico para los datos a usar, en la Figura N° 29 se muestra el formato usado en el presente proyecto.

```
@RELATION prestamos

@ATTRIBUTE monto NUMERIC
@ATTRIBUTE interes NUMERIC
@ATTRIBUTE ingreso NUMERIC
@ATTRIBUTE plazos NUMERIC
@ATTRIBUTE estadocivil NUMERIC
@ATTRIBUTE edad NUMERIC
@ATTRIBUTE class {no, yes}

@DATA
3000.0000,2.0000,1600.0000,2,1,62,yes
8000.0000,3.0000,600.0000,3,1,71,yes
3000.0000,1.8000,1000.0000,4,1,58,yes
3000.0000,1.8000,1200.0000,2,1,68,yes
17000.0000,2.0000,.0000,1,1,72,no
3500.0000,2.0000,.0000,1,1,60,no
```

Figura N° 29 Formato del archivo de datos para Weka
Fuente: Elaboración propia

4.4.3. Arquitectura de la aplicación

4.4.3.1. Componentes de la aplicación

Para el proyecto se han realizado tres componentes fundamentes.

a) Componente Web

Este componente contiene la página web para ingresar los datos que serán usados para clasificar el nuevo préstamo crediticio e indicar si ha sido rechazado o aprobado. Para lo cual se ha usado Python en la parte de la página web.

b) Componente API

Este componente contiene los métodos que serán consumidos por el sistema web. Los métodos implementados permiten calcular el nivel de exactitud de la clasificación de nuevos préstamos crediticios, indica si un préstamo será aceptado o rechazado según los datos recibidos. Se ha usado el microframework Flask para la implementación.

c) Componente ML

Este componente es el más importante y trabaja en el nivel más bajo de la aplicación, ya que por medio de este componente se realiza toda la lógica del algoritmo de machine learning que se implementará, además de consumir la información con la cual se entrenará el algoritmo. Se ha desarrollado con Python y numpy.

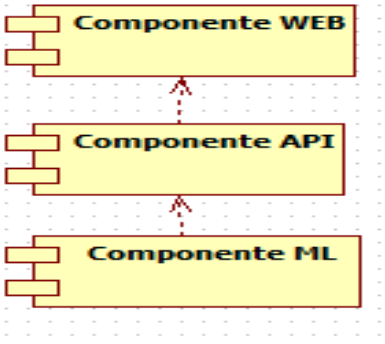


Figura N° 30 Componentes de la aplicación
Fuente: Elaboración propia

4.4.4. Construcción del modelo

El modelo de machine learning desarrollado es una regresión logística la cual ha sido construida con los siguientes algoritmos:

4.4.4.1. Algoritmo de escalamiento de datos

El algoritmo de escalamiento permite reducir el rango de dispersión de los datos y de esa manera el algoritmo de regresión logística puede hacer los cálculos de manera más rápida, además que ayuda a reducir el sobre ajuste del conjunto de datos de entrenamiento.

Se ha implementado de la siguiente manera:

Tabla N° 38 Fórmula de escalamiento de datos

Fórmula matemática	Pseudocódigo
$X_i = \frac{X_i - \mu_i}{S_i}$	$function\ Escalar(X : matriz)$: $Para\ i = 1\ hasta\ n$: $mu \leftarrow promedio(X_{(i)})$ $sigma \leftarrow destand(X_{(i)})$ $X_{(i)} \leftarrow \frac{(X_{(i)} - mu)}{sigma}$ end $retornar\ X$
Dónde: $S_i = desviación\ estándar$ $\mu_i = media\ de\ los\ valores$ $X = conjunto\ de\ datos$	

4.4.4.2. Función Sigmoidal

La función sigmoidal servirá para obtener los valores predictivos de nuevos préstamos crediticios y también para calcular el valor óptimo de clasificación de cada coeficiente de entrenamiento, entiéndase por coeficientes a los valores θ (theta) que se tendrán que calcular por medio de las demás funciones.

Tener en cuenta que la función sigmoidal usada para la regresión logística es uniporlar, es decir, los valores estarán en el rango $[0, 1]$.

Se ha implementado de la siguiente manera:

Tabla N° 39 Algoritmo para función sigmodal

Fórmula matemática	Pseudocódigo
$Sigmoid = \frac{1}{1 + e^{-\theta^T X}}$	$function\ Sigmd(X\ , theta: matriz)$: $Para\ i = 1\ hasta\ m$: $z_i \leftarrow\ X_i * \theta$ end $sig \leftarrow \frac{1}{(1 + e^{-z})}$ $retornar\ sig$
Dónde: $e = función\ exponencial$ $\theta = valores\ de\ los\ coeficientes$ $X = conjunto\ de\ datos$	

4.4.4.3. Función de costo para el modelo de Regresión Logística

La función de costo sirve para verificar que el modelo converja, es decir, minimizar en cada iteración para que se obtengan los valores de los coeficientes θ. Dado que se cuenta con dos clases $y \in \{0,1\}$ para clasificar, la función de costo se reduce a dos casos de evaluación de donde se deduce la siguiente fórmula para el modelo propuesto.

Tabla N° 40 Algoritmo para función de costo

Formulación matemática
$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^{(i)}) - y^{(i)})$ $Si\ y = 1, Cost(h_\theta(x^{(i)}) - y^{(i)}) = -\log(h_\theta(x^{(i)}))$ $Si\ y = 0; Cost(h_\theta(x^{(i)}) - y^{(i)}) = -\log\left(1 - h_\theta(x^{(i)})\right)$ *Finalmente se obtiene la función de costo* $J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} * \log\left(h_\theta(x^{(i)})\right) + (1 - y^{(i)}) * \log(1 - h_\theta(x^{(i)}))\right]$
Dónde: $m = cantidad\ de\ registros\ en\ el\ conjunto\ de\ entrenamiento$ $\theta = valores\ de\ los\ coeficientes\ para\ cada\ característica$ $y = valores\ objetivo\ o\ clases\ y \in \{0,1\}$ $h_\theta(x^{(i)}) = función\ sigmoidal\ o\ también\ llamada\ función\ de\ hipótesis$
Pseudocódigo
$function\ CostoLR(X, y, theta : matriz)$: $costoTotal \leftarrow 0$ $Para\ i = 1\ hasta\ m$: $xpos \leftarrow X_{(i)}$ $H \leftarrow Sigmd(theta, xpos)$ $Si\ y_{(i)} = 1\ entonces$: $costLocal \leftarrow y_{(i)} * \log(H)$ $Sino\ si\ y_{(i)} = 0\ entonces$: $costLocal \leftarrow (1 - y_{(i)}) * \log(1 - H)$ $costoTotal \leftarrow costoTotal + costLocal$ $value \leftarrow (-1/m)$ $J \leftarrow value * costoTotal$ $retornar\ J$

4.4.4.4. Función de la Gradiente de Descenso

Ya que se ha realizado la función de costo, ahora la función de la gradiente de descenso que es un algoritmo de optimización iterativo de primer orden, lo que permite encontrar los valores finales de θ que se usarán para predecir nuevas instancias de préstamos crediticios. Para realizar la gradiente de descenso se ha tomado como valor de aprendizaje $\alpha = 0.04$ y un total de 400 iteraciones.

Tabla N° 41 Algoritmo para gradiente de descenso

Formulación matemática
$min_\theta J(\theta)$: $Hasta\ converger,$ $$\theta_j = \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^i$$
Pseudocódigo
$function\ Gradiente(X\ , y\ , theta : matriz)$: $\quad newTheta \leftarrow theta$ $\quad Para\ j = 1\ hasta\ len(theta)$: $\quad\quad derivada \leftarrow CostoGradiente(X\ , y\ , newTheta)$ $\quad\quad newTheta \leftarrow newTheta_j - derivada$ $\quad retornar\ newTheta$ $function\ CostGradiente(X\ , y\ , theta : matriz\ , alpha : decimal)$: $\quad costTotal \leftarrow 0$ $\quad Para\ i = 1\ hasta\ m$: $\quad\quad xpos \leftarrow X_{(i)}$ $\quad\quad xposij \leftarrow xpos_j$ $\quad\quad H \leftarrow Sigmd(xpos\ , theta)$ $\quad\quad costoLocal \leftarrow \left(H - y_{(i)}\right) * xposij$ $\quad\quad costoTotal \leftarrow costoTotal + costoLocal$ $\quad value \leftarrow (alpha/m)$ $\quad G = value * costoTotal$ $\quad retornar\ G$

4.5. Interfaz de usuario de la aplicación

En la siguiente imagen se puede observar la interfaz de la aplicación, la cual tiene un formulario con características necesarias para que el algoritmo de machine learning realice la clasificación. Además, es de fácil uso para el usuario.

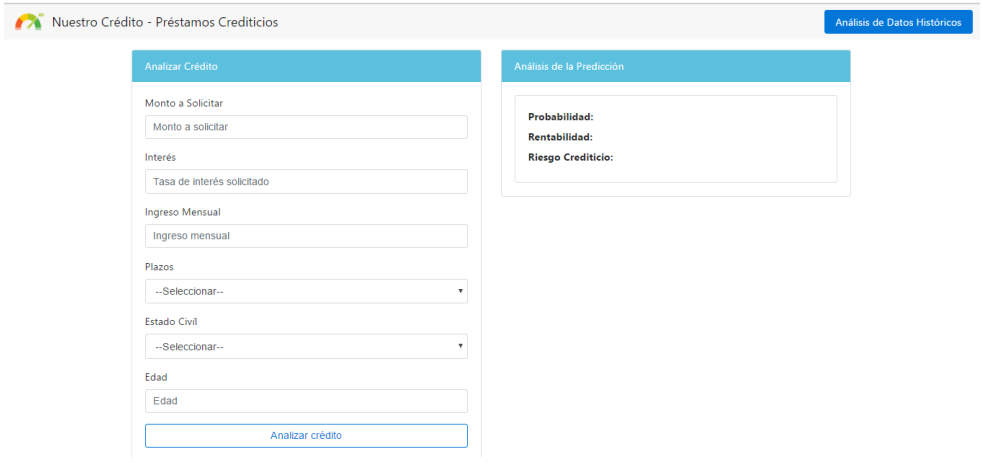


Figura N° 31 Interfaz de para predicción préstamos crediticios
Fuente: Elaboración propia

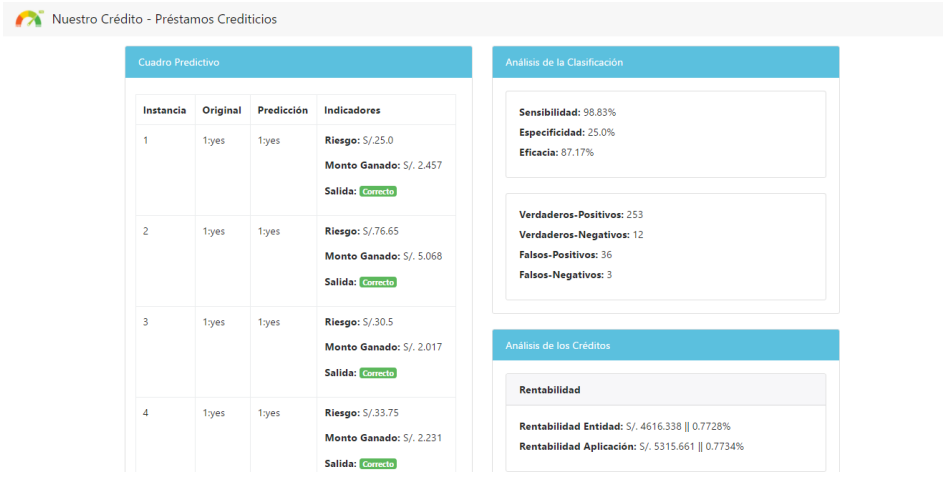


Figura N° 32 Interfaz de análisis de datos históricos
Fuente: Elaboración propia

4.6. Pruebas

Las pruebas realizadas al modelo de regresión logística que se han desarrollado en el presente proyecto fueron hechas en la herramienta Weka, la cual brinda la información de los préstamos crediticios clasificadas para el conjunto de datos pruebas (25%).

Los pasos para realizar las pruebas del modelo fueron las siguientes:

a) Seleccionar el archivo en formato (.arff) que contiene el conjunto de datos que se va a utilizar en las pruebas.

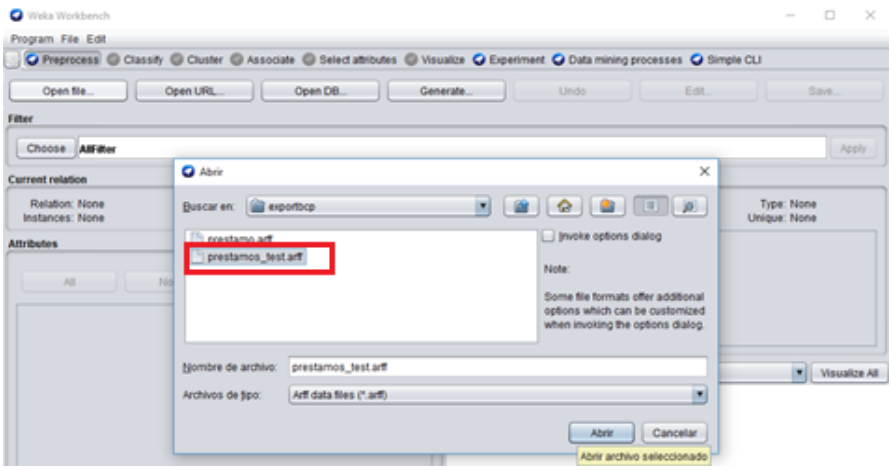


Figura N° 33 Selección de archivo para Weka
Fuente: Elaboración propia

b) Indicamos el modelo a usar, en este caso Regresión Logística, y luego seleccionamos el valor (%) que se usará para entrenar el modelo y el valor restante (%) se usará para las pruebas.

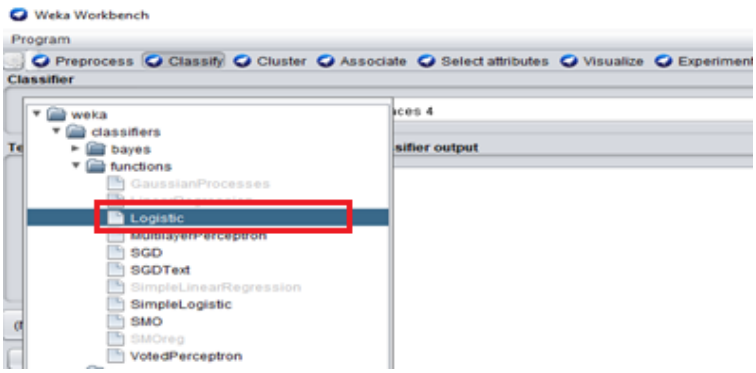


Figura N° 34 Selección de modelo de aprendizaje
Fuente: Elaboración propia

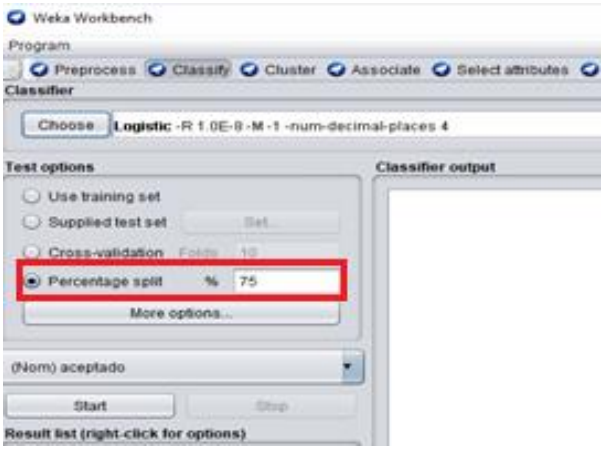


Figura N° 35 Indicador de porcentaje de partición de datos
Fuente: Elaboración propia

c) Finalmente se realizó la prueba al modelo seleccionado y se obtuvieron los siguientes resultados a una precisión de 88.1579% para el 25% de los datos:

Tabla N° 42 Resultados de clasificación de instancias

Instancia	Valor actual	Valor de predicción	Error de predicción
1	2:yes	2:yes	0.956
2	2:yes	2:yes	0.506
3	1:no	2:yes (+)	0.925
4	2:yes	2:yes	0.981
5	2:yes	2:yes	0.965

Para ver la tabla completa de los resultados ver el Anexo 02.

Tabla N° 43 Matriz de confusión

		Clasificado como	
		YES	**NO**
Valor de predicción	**YES**	66	1
	NO	8	1

4.7. Implantación

4.7.1. Planificar la implantación

4.7.1.1. Descripción de resultados, modelo y descubrimientos

Los resultados obtenidos luego de realizar pruebas sobre el modelo de machine learning – Regresión Logística – se obtuvo una precisión de 88.1594% al 25% del conjunto de datos, de esta manera se valida que el modelo seleccionado para el presente proyecto tiene un nivel de predicción acorde con lo que se requiere según los objetivos del negocio.

El modelo de Regresión Logística no es un modelo complejo pero ayuda al cumplimiento de los objetivos del proyecto, por lo cual fue seleccionado. Este modelo está basado en clasificación de clases (sí, no), además se ha incluido el monto del riesgo de cada préstamo crediticio.

Los descubrimientos encontrados están en los patrones entre los montos requeridos, la edad de las personas y sus ingresos. En las siguientes gráficas se muestran los patrones de los datos.

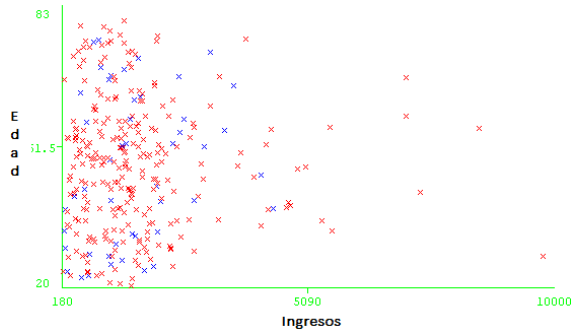


Figura N° 36 Patrones de ingreso vs edad
Fuente: Elaboración propia

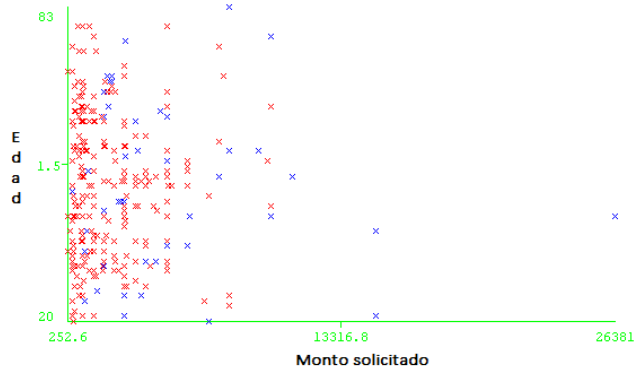


Figura N° 37 Patrones de monto solicitado vs edad
Fuente: Elaboración propia

En Figura N° 36 se observan patrones que indican que la mayor cantidad de ingresos están en las personas con edades entre los 20 a 80 años de edad y en la Figura N° 37 se observan patrones de los montos que se solicitan en su mayoría son menos de 10,000 soles y el rango de edad de 20 a 80 años.

4.7.2. Despliegue de la aplicación

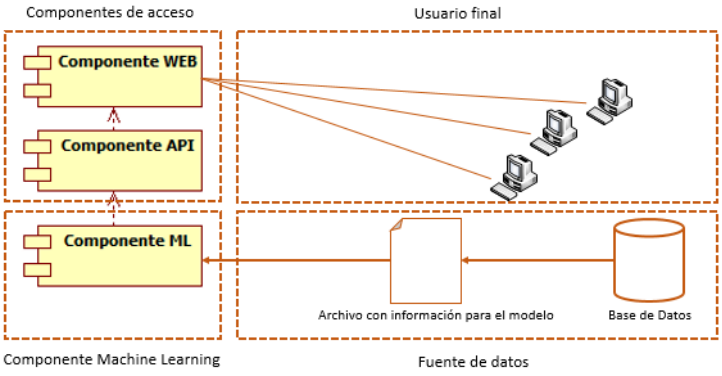


Figura N° 38 Diagrama de despliegue
Fuente: Elaboración propia

CAPÍTULO 5. METODOLOGÍA

5.1. Diseño de investigación

Investigación pre experimental, porque se pretende investigar la forma en que el desarrollo de una aplicación basada en un modelo de machine learning mejora la evaluación de préstamos crediticios.

Se cuenta con un grupo “G1”, la variable a evaluar “X” y dos observaciones “O1” y “O2”

G1	X	O1
		O2

5.2. Unidad de estudio

Información de un préstamo crediticio.

5.3. Población

La clasificación la conforman 1460 préstamos crediticios realizados por la financiera Nuestro Crédito comprendidos en el periodo de enero hasta diciembre de los años 2014 y 2015.

5.4. Muestra

La muestra la conforman 304 préstamos crediticios realizados por la financiera Nuestro Crédito comprendidos en el periodo de enero hasta diciembre de los años 2014 y 2015.

$$n = \frac{N\,\sigma^2\,Z^2}{e^2\,(N-1) + \sigma^2 Z^2}$$

$$n = \frac{1450\,(0.5)^2\,(1.96)^2}{(0.05)^2\,(1450-1) + (0.5)^2(1.96)^2} = 304$$

5.5. Técnicas, instrumentos y procedimientos de recolección de datos

- **Entrevista:** Esta técnica se usa para la recolección de datos en la realidad problemática.
- **Observación:** Esta técnica se usa en la evaluación de la variable dependiente e independiente.
- **Hojas de verificación:** Este instrumento se usa para registrar los datos obtenidos del análisis de los indicadores.

Tabla Nº 44 Técnica e instrumentos de la variable dependiente

Variable dependiente	Dimensión	Indicador	Técnica	Instrumento

Evaluación de préstamos crediticios	Rentabilidad	Porcentaje de dinero ganado por los préstamos crediticios clasificados	Observación	Hoja de verificación
	Riesgo crediticio	Porcentaje de dinero perdido por préstamos crediticios clasificados.	Observación	Hoja de verificación
	Tiempo	Tiempo promedio en días para aprobar un préstamo crediticio.	Observación	Hoja de verificación

Fuente: Elaboración propia

Tabla Nº 45 Técnicas e instrumentos de la variable independiente

Variable independiente	Dimensión	Indicador	Técnica	Instrumento
Aplicación informática basada en un modelo de machine learning	Sensibilidad	Porcentaje de préstamos crediticios positivas clasificadas como aprobados	Observación	Hoja de verificación
	Especificidad	Porcentaje de préstamos crediticios clasificados como rechazados	Observación	Hoja de verificación
	Eficacia	Porcentaje de préstamos crediticios	Observación	Hoja de verificación

		clasificados de manera correcta		

Fuente: Elaboración propia

5.6. Métodos, instrumentos y procedimientos de análisis de datos

- Realizar investigación sobre modelos de machine learning y seleccionar el modelo que mejor se adapte al desarrollo para la evaluación de préstamos crediticios.
- Luego de analizar, seleccionar las características que servirán para construir el modelo de evaluación de préstamos crediticios; teniendo en cuenta la información con la que se cuenta.
- Identificar un conjunto de datos para realizar la fase de entrenamiento del modelo tomando el 75% de la muestra y el 25% para pruebas del modelo a implementar.
- Realizar la medición de las variables por medio de las fórmulas que se plantean y dar los resultados obtenidos.

Tabla Nº 46 Métodos y procedimientos de la variable dependiente

Variable dependiente	Dimensión	Indicador	Método	Procedimiento
Evaluación de préstamos crediticios	Rentabilidad	Porcentaje de dinero ganado por los préstamos crediticios clasificados	$\frac{\sum GN}{\sum IV} \times 100$ ***GN:*** Ganancia de cada préstamo ***IV:*** Inversión de cada préstamo	Hoja de verificación
	Riesgo crediticio	Cantidad de dinero perdido por préstamos crediticios clasificados.	$\sum_{i=1}^{r} PD * EAD * (1 - R)$ ***PD:*** Probabilidad por defecto. ***EAD:*** Exposición al defecto. ***R:*** Tasa de recuperación.	Hoja de verificación

			r: Total de préstamos clasificados de manera correcta.	
	Tiempo	Tiempo promedio en días para aprobar un préstamo crediticio	$\frac{\sum_{i=1}^{r} TA - TP}{r}$ ***TP:*** Tiempo en el que se solicitó el préstamo. ***TA:*** Tiempo en ser aprobado el préstamo. ***r:*** Total de préstamos aprobados de manera correcta.	Hoja de verificación

Fuente: Elaboración propia

Tabla Nº 47 Métodos y procedimientos de la variable independiente

Variable independiente	Dimensión	Indicador	Método	Procedimiento
Aplicación informática basada en un modelo de machine learning	Sensibilidad	Porcentaje de préstamos crediticios clasificados como aprobados	$\frac{VP}{VP + FN} * 100$ ***VP:*** Verdaderos positivos ***FN:*** Falsos negativos	Hoja de verificación
	Especificidad	Porcentaje de préstamos crediticios clasificados como rechazados	$\frac{VN}{VN + FP} * 100$ ***VN:*** Verdaderos negativos ***FP:*** Falsos positivos	Hoja de verificación
		Porcentaje de préstamos	$\frac{VP + VN}{TI} * 100$	

	Eficacia	crediticios clasificadas de manera correcta	***TI:*** Total de instancias ***VP:*** Verdades positivos **VN:** Verdades negativos.	Hoja de verificación

Fuente: Elaboración propia

CAPÍTULO 6. RESULTADOS

A continuación, se muestran los resultados obtenidos de cada uno de los indicadores para la variable independiente y dependiente del proyecto de investigación, los cual contribuye en la contrastación de la hipótesis.

Tabla Nº 48 Resultados de clasificación

Préstamo crediticio	Valor actual	Valor de predicción
1	yes	yes
2	yes	yes
3	yes	yes
4	yes	no
5	no	yes
6	no	no
7	no	no
8	no	yes
9	no	yes
10	no	yes

En la tabla Nº 48 se han elegido aleatoriamente algunos de los resultados, para ver la tabla completa de los resultados de la muestra, ver el Anexo 03.

Tabla Nº 49 Matriz de confusión

		Valor actual del préstamo crediticio	
		YES	NO
Valor de predicción	YES	253	36
	NO	3	12

6.1. Resultados para los indicadores de la variable independiente

6.1.1. Indicador 1: Porcentaje de préstamos crediticios clasificados como aprobados

Tabla Nº 50 Descripción del primer indicador de la variable independiente

Nº	Indicador	Descripción del indicador	Unidad de medida	Instrumentos	Fórmula

01	Porcentaje de préstamos crediticios clasificados como aprobados	Porcentaje de préstamos crediticios que la aplicación informática clasificó como aprobados	Porcentaje	Hoja de verificación	$\frac{VP}{VP + FN} * 100$ ***VP:*** Verdaderos positivos ***FN:*** Falsos negativos

Considerando los datos de la Tabla Nº 49, representamos en la fórmula los datos correspondientes.

$$\frac{VP}{VP + FN} * 100$$

$$\frac{253}{253 + 3} * 100$$

98.83 % , de sensibilidad

6.1.2. Indicador 2: Porcentaje de préstamos crediticios clasificados como rechazados

Tabla Nº 51 Descripción del segundo indicador de la variable independiente

Nº	Indicador	Descripción del indicador	Unidad de medida	Instrumento	Fórmula
02	Porcentaje de préstamos crediticios clasificados como rechazados	Porcentaje de préstamos crediticios que la aplicación informática clasificó como rechazados	Porcentaje	Hoja de verificación	$\frac{VN}{VN + FP} * 100$ ***VN:*** Verdaderos negativos ***FP:*** Falsos positivos

Asignamos los valores en la fórmula, según los datos de la Tabla Nº 49.

$$\frac{VN}{VN + FP} * 100$$

$$\frac{12}{12 + 36} * 100$$

25%, de especificidad

6.1.3. Indicador 3: Porcentaje de préstamos crediticios clasificados de manera correcta

Tabla Nº 52 Descripción del tercer indicador de la variable independiente

Nº	Indicador	Descripción del indicador	Unidad de medida	Instrumento	Fórmula
03	Porcentaje de préstamos crediticios clasificados de manera correcta	Porcentaje de préstamos crediticios que la aplicación informática clasificó de manera correcta	Porcentaje	Hoja de verificación	$\frac{VP + VN}{TI} * 100$
					TI: Total de préstamos ***VP:*** Verdades positivos **VN:** Verdades negativos.

Representamos la fórmula con los valores indicados en la Tabla Nº 49.

$$\frac{VP + VN}{TI} * 100$$

$$\frac{253 + 12}{253 + 3 + 12 + 36} * 100$$

87,17%, de eficacia.

6.2. Resultados para los indicadores de la variable dependiente

6.2.1. Indicador 4: Porcentaje de dinero ganado por lo préstamos crediticios clasificados

Tabla Nº 53 Descripción del primer indicador de la variable dependiente

Nº	Indicador	Descripción del indicador	Unidad de medida	Instrumento	Fórmula
04	Porcentaje de dinero ganado por préstamos crediticios clasificados	Porcentaje de dinero ganado por los préstamos clasificados como aprobados	Porcentaje	Hoja de verificación	$\frac{\sum GN}{\sum IV} \times 100$
					GN: Ganancia de cada préstamo ***IV:*** Inversión de cada préstamo

Para la determinación del indicador, se calculó la rentabilidad de los préstamos crediticios clasificados por la entidad financiera y los préstamos predichos por la aplicación informática. Primero se calcula la ganancia de cada préstamo clasificado, a través de las siguientes fórmulas:

a. Calcular la tasa efectiva mensual:

$$tem = (1 + tasa\ de\ interes)^{1/12} - 1$$

b. Calcular el pago mensual:

$$pago\ mensual = \frac{monto * (tem + 1)^{cuotas}}{(1 + tem)^{cuotas} - 1}$$

c. Calcular el pago total:

$$pago\ total = pago\ mensual * cuotas$$

d. Calcular la ganancia:

$$ganancia = pago\ total - monto$$

Tabla Nº 54 Resultados de los montos ganados

Instancia	Valor actual	Monto actual	Valor de predicción	Monto ganado
247	yes	24.807	yes	24.807
248	yes	8.629	yes	8.629
249	yes	16.52	yes	16.52
250	yes	43.425	no	0.0
251	yes	2.478	yes	2.478
252	yes	5.414	yes	5.414
253	yes	9.2	yes	9.2
254	yes	10.834	yes	10.834
255	yes	63.084	yes	63.084
247	yes	24.807	yes	24.807

En la tabla Nº 54 se han elegido aleatoriamente algunos de los resultados. Para ver la tabla completa de los resultados de la muestra, ver el Anexo 04.

- Rentabilidad de la clasificación de la entidad financiera,

$$RF = \frac{\sum GNF}{\sum IVF} * 100$$

$$RF = \frac{4616.338}{597324.6} * 100$$

$$RF = 0.7728\ \%$$

- Rentabilidad de la clasificación de la aplicación informática.

$$RA = \frac{\sum GNA}{\sum IVA} * 100$$

$$RF = \frac{5315.661}{687326.6} * 100$$

$$RF = 0.7734\%$$

La diferencia de rentabilidad es de 0.0006 %, lo cual indica que al utilizar la aplicación informática la entidad financiera gana S/. 699.323 más que con el método actual de evaluación de préstamos crediticios.

6.2.2. Indicador 5: Cantidad de dinero perdido por préstamos crediticios clasificados

Tabla Nº 55 Descripción del segundo indicador de la variable dependiente

Nº	Indicador	Descripción del indicador	Unidad de medida	Instrumentos	Fórmula
05	Cantidad de dinero perdido por préstamos crediticios clasificados	Monto de dinero que se ha perdido por préstamos clasificados	Soles	Hoja de verificación	$\sum_{i=1}^{r} PD * EAD * (1 - R)$ ***PD:*** Probabilidad por defecto. ***EAD:*** Exposición al defecto. ***R:*** Tasa de recuperación ***r:*** Total de préstamos clasificados de manera correcta.

Para la determinación del indicador, se calculó el riesgo crediticio de los préstamos clasificados por la entidad financiera y los préstamos predichos por la aplicación informática.

Tabla Nº 56 Resultados de riesgo

Instancia	Valor actual	Riesgo actual	Valor de predicción	Riesgo de predicción
34	yes	375.0	yes	375.0
35	yes	115.0	yes	115.0
36	yes	100.0	yes	100.0
37	yes	30.0	yes	30.0
38	yes	385.0	no	0.0
39	yes	250.0	yes	250.0
40	yes	175.0	yes	175.0
41	yes	19.0	yes	19.0
42	yes	19.0	yes	19.0
34	yes	375.0	yes	375.0

En la tabla Nº 56 se han elegido aleatoriamente algunos de los resultados; para ver la tabla completa de los resultados de la muestra, ver el Anexo 05.

Para la aplicación de la fórmula se reemplaza los siguientes valores:

a) Probabilidad de default (PD): 0.10

b) Exposición a default (EAD) = monto solicitado

c) Pérdida en caso de incumplimiento = (1 – R) = 0.40

- Riesgo crediticio de los préstamos clasificados por la entidad financiera

$$\sum_{i=1}^{r} PD * EAD * (1 - R)$$

S/. 29866.23

- Riesgo crediticio de los préstamos predichos por la aplicación informática.

$$\sum_{i=1}^{r} PD * EAD * (1 - R)$$

S/. 28616.23

Al obtener el riesgo crediticio, se puede apreciar que el valor obtenido por los préstamos crediticios clasificados por la aplicación informática es menor que el valor obtenido por los préstamos clasificados de la entidad financiera. Disminuyendo en un valor de S/. 1250 el riesgo total de los préstamos crediticios aceptados.

6.2.3. Indicador 6: Tiempo promedio en días para aprobar un préstamo crediticio

Tabla Nº 57 Descripción del tercer indicador de la variable dependiente

Nº	Indicador	Descripción del indicador	Unidad de medida	Instrumentos	Fórmula
06	Tiempo promedio en días para aprobar un préstamo crediticio.	Tiempo en días utilizado para aprobar un préstamo.	Tiempo	Hoja de verificación	$\frac{\sum_{i=1}^{r} TA - TP}{r}$ ***TP:*** Tiempo en el que se solicitó el préstamo. ***TA:*** Tiempo en ser aprobado el préstamo. ***r:*** Total de préstamos aprobados de manera correcta.

Para calcular este indicador, se consideró el tiempo de evaluación de los préstamos clasificados por la entidad y los préstamos predichos por la aplicación, este último se calcula desde que el usuario hace clic en el botón Analizar crédito.

Tabla Nº 58 Resultados de tiempo

Instancia	Valor actual	Tiempo actual (Segundos)	Valor de predicción	Tiempo de predicción (Segundos)
34	yes	259200	yes	0.2633
35	yes	172800	yes	1.0007
36	yes	172800	yes	0.2388
37	yes	259200	yes	0.2269
38	yes	172800	yes	0.2209
164	no	86400	yes	1.0051
165	no	259200	yes	0.7559
166	no	86400	no	1.0513
167	no	86400	no	0.5712
168	no	86400	yes	0.7644

En la tabla Nº 58 se han seleccionado aleatoriamente algunos de los resultados. Para ver la tabla completa de los resultados de la muestra, ver el Anexo 06.

- Tiempo promedio de los préstamos clasificados por la entidad financiera

$$\frac{\sum_{i=1}^{r} TA - TP}{r}$$

$$2.03\ días$$

- Tiempo promedio de los préstamos predichos por la aplicación informática

$$\frac{\sum_{i=1}^{r} TA - TP}{r}$$

$$8e^{-06} = 0.00000796\ días$$

Al obtener el tiempo promedio, se puede apreciar que existe una gran diferencia entre los tiempos de evaluación de los préstamos crediticios por parte de la entidad financiera y la aplicación informática desarrollada. Por eso, se concluye que el tiempo promedio de la solución planteada es mucho más rápida.

CAPÍTULO 7. DISCUSIÓN

Esta investigación tiene como propósito determinar la mejora de la evaluación de préstamos crediticios mediante una aplicación informática basada en un modelo de machine learning.

Después de haber procesado los indicadores de la variable independiente, se obtuvo lo siguiente:

Indicador 1: El porcentaje de préstamos crediticios clasificados como aprobados es 98.93%, lo que significa que la aplicación informática clasifica de manera correcta 253 de los 256 préstamos crediticios aprobados que se utilizaron en la muestra. Por eso, el porcentaje de sensibilidad es aceptable para la evaluación de nuevos préstamos crediticios con el uso de la aplicación informática desarrollada.

Indicador 2: El porcentaje de préstamos crediticios clasificados como rechazados es 25%, lo que significa que la aplicación informática clasifica 12 de los 48 préstamos crediticios rechazados que se utilizó en la muestra. Esto indica que el porcentaje de especificidad es bajo, es decir, que hay un 75% de préstamos crediticios rechazados que fueron clasificados como préstamos crediticios aprobados por medio de la aplicación informática. Aunque el porcentaje de especificidad es bajo, esto no indica que el modelo desarrollado es incorrecto, ya que esto depende del porcentaje de eficacia que se obtenga en la evaluación total.

Indicador 3: El porcentaje de préstamos crediticios clasificados de manera correcta es 87.17%, lo que indica que la eficacia del modelo de machine learning seleccionado es adecuado, pues se obtuvo un resultado de evaluación mayor al 80% según el criterio de éxito del negocio. Además en comparación con los resultados de los antecedentes, el algoritmo de regresión logística realiza de manera más eficaz la clasificación de los préstamos crediticios que los algoritmos j48 con un valor de 78.38%, redes bayesianas con 77.47% y bayes ingenuo 77.87% los cuales son planteados por Aboobyda & Tarig; la mejora obtenida es de 8.79%, 9.7% y 9.3% respectivamente. También se obtiene una mejora de 2.17% en comparación con el algoritmo de máquina de soporte de vectores planteado por Kim A. & Lo. En cuanto a la comparación con el uso del modelo K-meas que pertenece a un modelo de aprendizaje no supervisado planteado por Kavitha K. se obtiene una mejora de 7.17%.

Luego de haber procesado los indicadores de la variable dependiente, se obtuvo lo siguiente:

Indicador 4: El porcentaje de dinero ganado por préstamos crediticios clasificados como aprobados por medio de la aplicación informática se obtuvo 0.7728% a comparación de lo obtenido por la entidad financiera con 0.7734%, dando una diferencia de 0.0006% lo que equivale a un aumento en la rentabilidad de S/. 699.323. El aumento de la rentabilidad es mínimo debido a que los intereses de los préstamos usados en la investigación varían entre el 1% y 4%. Con estos datos obtenidos se demuestra que la aplicación informática mejora el incremento del porcentaje de dinero ganado.

Indicador 5: La cantidad de dinero perdido por los préstamos crediticios clasificados como aprobados por medio de la aplicación informática es S/. 28616.23 en comparación con lo obtenido por la entidad financiera que es S/. 29866.23, se obtiene una disminución del riesgo de S/. 1250 equivalente al 4.19%. Con estos datos obtenidos se demuestra que la aplicación informática disminuye la cantidad de dinero perdido al evaluar un préstamo crediticio.

Indicador 6: El tiempo promedio en días para aprobar un préstamo crediticio por medio de la aplicación informática es de 0.00000796 días lo que equivale a 1.6 segundos, en comparación con el tiempo promedio que se toma la entidad financiera que es de 2 días. Logrando una mejora en el tiempo de aprobación de préstamo crediticios del 99.9%.

CONCLUSIONES

Al finalizar la investigación de este proyecto, se obtuvieron las siguientes conclusiones:

- Se logró mejorar la evaluación de préstamos crediticios mediante una aplicación informática basada en un modelo de machine learning, cumpliendo los siguientes objetivos específicos.
- Se logró aumentar el porcentaje de dinero ganado por la clasificación de préstamos crediticios en un 0.0006% lo que equivale a S/. 699.323 más que la forma actual de evaluación de los préstamos crediticios.
- Se logró disminuir la cantidad de dinero perdido por la clasificación de préstamos crediticios de S/. 29866.33 a S/. 28616.23 lo que equivale a un 4.19%.
- Se logró disminuir el tiempo promedio en días para aprobar un préstamo crediticio de 2 días a 0.00000796 días; es decir a 1.6 segundos que equivale al 99.9%. Este tiempo promedio es lo que necesita la aplicación informática para evaluar de manera eficiente la solicitud de un préstamo crediticio.
- Se logró aumentar el porcentaje de préstamos crediticios de manera correcta con una eficiencia de 87.17% por medio de la aplicación informática con el modelo de regresión logística.

RECOMENDACIONES

Con la finalización de esta investigación, se recomienda considerar los siguientes puntos:

- Analizar el impacto de una mejora para el algoritmo de optimización de la gradiente de descenso planteado en la fase de construcción para obtener una mejor eficacia del modelo de machine learning.
- Considerar en la etapa de pruebas distintos algoritmos de clasificación como el algoritmo de árbol de decisión para mejor el nivel de predicción, por lo cual es necesario contar con una cantidad mayor de datos de entrenamiento.
- Evaluar y agregar nuevas características de acuerdo a la empresa financiera donde se implemente el modelo de regresión logística planteado en la investigación, ya que estas pueden variar.

REFERENCIAS

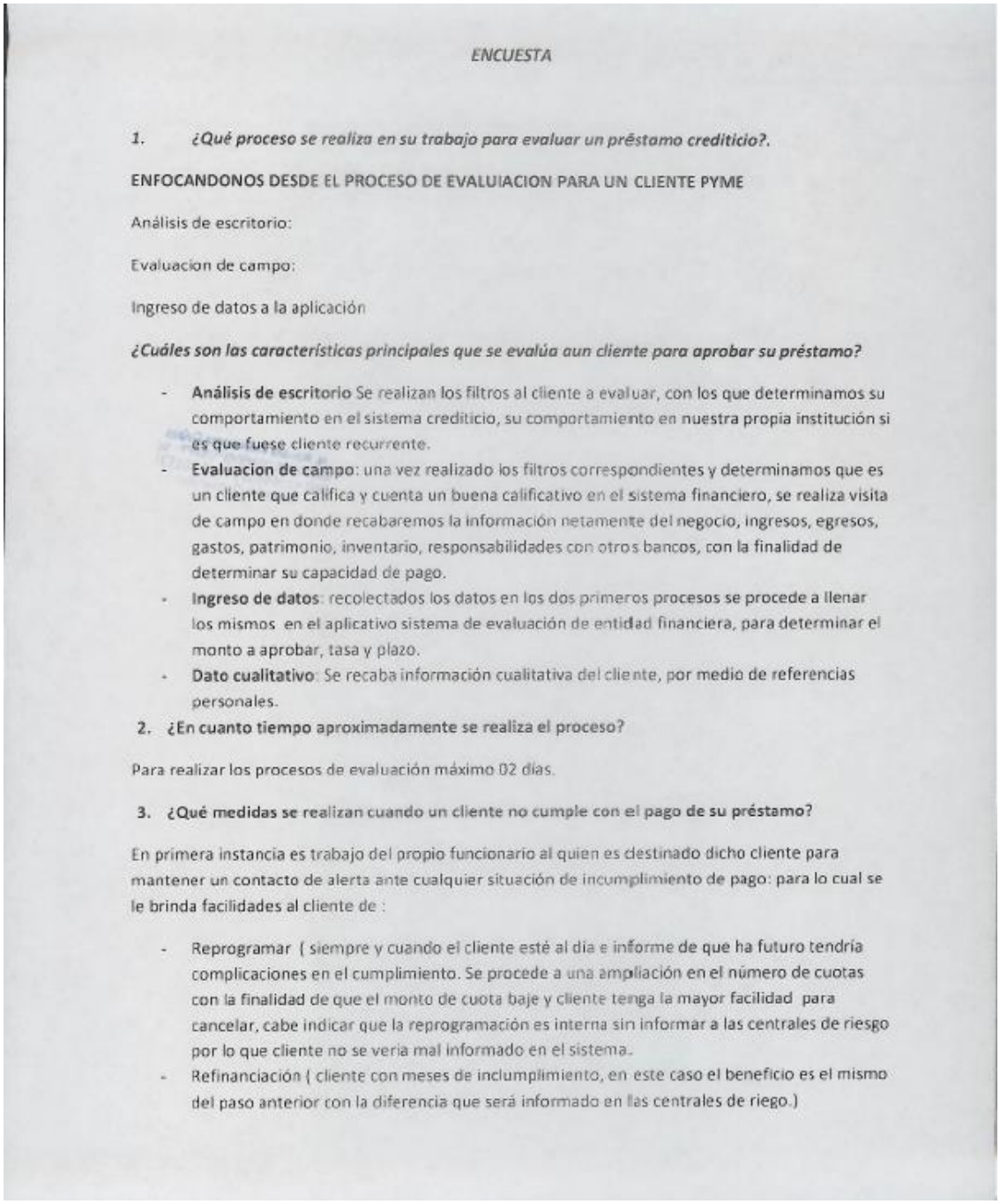
- Aboobyda J. & Taring M. (2016). *Developing prediction model of loan risk in banks using data mining (Vol. 3).* Khartoum, Sudan.
- Alpaydin E. (2010). *Introduction to Machine Learning (2da Edición)*. Cambridge, Estados Unidos: Massachusetts Institute of Technology.
- Añez Manfredo (24 de junio del 2001). *Aspectos básicos del análisis de créditos*. Recuperado de https://www.gestiopolis.com/aspectos-basicos-del-analisis-de-credito/
- Banco Central de Reserva del Perú. (Mayo del 2009). *IV Concurso escolar BCRP.* Lima: BCRP.
- Barrientos Martínez, R., Cruz Ramírez, N., Acosta Meza, H., Rabatte Suarez, I., otros. (Enero-junio 2008) *Evaluación del Potencial de Redes Bayesianas en la Clasificación en Datos Médicos. Revista Médica de la Universidad Veracruzana.* Vol 8, num. 1, pg. 33-37.
- BBVA (2017). *Préstamo Simple*. Recuperado de https://www.bbvacontinental.pe/personas/prestamos/personales/prestamo-libre-disponibilidad/
- BBVA (26 de octubre del 2016). *Diferencias entre un préstamo y una línea de crédito*. Recuperado de https://www.bbva.com/es/noticias/economia/bancos/diferencias-entre-un-prestamo-y-un-credito/
- BBVAOPEN4U (1 de septiembre del 2015). *¿Necesitas un préstamo? Tu única posibilidad es convencer a un algoritmo*. Recuperado de https://bbvaopen4u.com/es/actualidad/necesitas-un-prestamo-tu-unica-posibilidad-es-convencer-un-algoritmo
- BCRP (mayo del 2009). *Importancia del crédito*. IV Concurso Escolar de BCRP.
- Betancourt, Gustavo (2005). *Las máquinas de soporte vectorial (SVMs). Scientia et Technica Año XI, N.* Nº 27, pg. 67-72.
- Chodorow K. (2013). *MongoDB The definitive guide.* (2da Edición). Estados Unidos: O'Reilly Media.
- Comercio (17 de octubre del 2014). *100 mil personas son excluidas del sistema financiero por año*. Recuperado de http://elcomercio.pe/economia/peru/100-mil-personas-son-excluidas-sistema-financiero-ano-noticia-1764619
- Dans E. (13 de julio del 2016). *La medida de la importancia de la inteligencia artificial*. Recuperado de https://www.enriquedans.com/2016/07/la-medida-de-la-importancia-de-la-inteligencia-artificial.html

- Galán Cortina, Víctor (22 de enero del 2016). *Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario.* Recuperado de http://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf?sequence=1
- Gómez Terejina (16 de abril del 2016). *El uso de machine learning en las entidades financieras.* Recuperado de https://blogs.deusto.es/bigdata/el-uso-del-machine-learning-en-las-entidades-financieras/
- Gonzales, Andres (2014). *¿Qué es Machine Learning?* Recuperado de http://cleverdata.io/que-es-machine-learning-big-data/
- Ian H. & Eibe F. (2005). *Data Mining Practical Machine Learning Tools and Techniques (2da Edición).* San Francisco, Estados Unidos: Elsevier.
- IBM (2012). *Manual CRISP-DM de IBM SPSS Modeler* 15.Estados Unidos: IBM Corporation
- Kavitha K. (2016). *Clustering loan applicants based on risk percentage using K-means clustering techniques (Vol. 3)*. Kodaikanal, India.
- Kim A & Lo A. (2010). *Consumer credit risk models via machine learning algorithms*. Cambridge, Estados Unidos.
- Konovalenko, M. (30 de julio del 2013). *Artificial Intelligence is the most important technology of the future*. Recuperado de https://mariakonovalenko.wordpress.com/2013/07/30/artificial-intelligence-is-the-most-important-technology-of-the-future/
- Llorca E. (2 de julio del 2012). *7 riesgos de los créditos personales y cómo superarlos*. Recuperado de http://www.iahorro.com/ahorro/gestiona_tus_finanzas/7-riesgos-de-los-creditos-personales-y-como-superarlos.html
- Malca S. (2015). *Modelo algorítmico para la clasificación de una hoja de planta en base a sus características de forma y textura.* Lima, Perú.
- Marr, Bernard (2016). *A Short History of Machine Learning - Every Manager Should Read.* Recuperado de https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/#4bc1af7a15e7
- Microsoft (2016). *Cómo elegir algoritmos para Aprendizaje automático de Microsoft Azure.* Recuperado de https://docs.microsoft.com/es-es/azure/machine-learning/machine-learning-algorithm-choice
- OPENBBVA4U (19 de julio del 2016). *Inteligencia y Artificial y Big Data aplicados al negocio bancario.* Recuperado de https://bbvaopen4u.com/es/actualidad/inteligencia-artificial-y-big-data-aplicados-al-negocio-bancario

- Oracle Corportation (2016). *Netbeans IDE*. Recuperado de http://www.oracle.com/technetwork/developer-tools/netbeans/overview/index.html
- Oracle Corpotation (2016). *¿Qué es la tecnología Java y para qué lo necesito?* California: Oracle Web. Recuperado de https://www.java.com/es/download/faq/whatis_java.xml
- Quispesaravia R. & Perez W (2015). *Herramienta de análisis y clasificación de complejidad de textos en español.* Lima, Perú.
- Robio Alex, Noviembre 2016. *Start preparing now for the emergence of Machine Learning.* Recuperado de http://blog.belatrixsf.com/start-preparing-now-for-the-emergence-of-machine-learning/
- Rouse M. (Enero del 2015). *SQL Server.* Recuperado de http://searchdatacenter.techtarget.com/es/definicion/SQL-Server
- Russell S. & Norvig P. (2004). *Inteligencia Artificial un enfoque moderno (2da Edición)*. España: Pearson Educación.
- Sánchez Morales (s.f). *Máquinas de aprendizaje extremo multicapa: Estudio y evaluación de resultados en la segmentación automática de carótidas en imágenes ecográficas (Proyecto)*. Universidad Politécnica de Cartagena.
- Smola y Vishwanathan (2008). *Introduction to Machine Learning*. United Kingdom: Cambridge University Press.
- Statistical Analysis System (2016). *Machine Learning what it is and what it matters.* Recuperado de https://www.sas.com/en_us/insights/analytics/machine-learning.html
- Tirados M. (15 de abril del 214). *Apache Spark, la nueva estrella del Big Data*. Recuperado de http://www.bigdatahispano.org/noticias/apache-spark-la-nueva-estrella-de-big-data/
- Vincec T. (07 de diciembre del 2011). *La Inteligencia Artificial*. Recuperado de http://www.fgcsic.es/lychnos/es_es/articulos/inteligencia_artificial
- Vindi N. (1 de septiembre del 2012). *Importancia de los préstamos en los bancos.* Recuperado de http://www.bolsamania.com/mejoresprestamos/importancia-de-los-prestamos-en-los-bancos/

ANEXOS

ANEXO 01 – ENTREVISTA DE PROCESO DE EVALUACIÓN DE PRÉSTAMOS CREDITICIOS



ENCUESTA

1. ***¿Qué proceso se realiza en su trabajo para evaluar un préstamo crediticio?.***

ENFOCANDONOS DESDE EL PROCESO DE EVALUIACION PARA UN CLIENTE PYME

Análisis de escritorio:

Evaluacion de campo:

Ingreso de datos a la aplicación

¿Cuáles son las características principales que se evalúa aun cliente para aprobar su préstamo?

- **Análisis de escritorio** Se realizan los filtros al cliente a evaluar, con los que determinamos su comportamiento en el sistema crediticio, su comportamiento en nuestra propia institución si es que fuese cliente recurrente.
- **Evaluacion de campo**: una vez realizado los filtros correspondientes y determinamos que es un cliente que califica y cuenta un buena calificativo en el sistema financiero, se realiza visita de campo en donde recabaremos la información netamente del negocio, ingresos, egresos, gastos, patrimonio, inventario, responsabilidades con otros bancos, con la finalidad de determinar su capacidad de pago.
- **Ingreso de datos**: recolectados los datos en los dos primeros procesos se procede a llenar los mismos en el aplicativo sistema de evaluación de entidad financiera, para determinar el monto a aprobar, tasa y plazo.
- **Dato cualitativo**: Se recaba información cualitativa del cliente, por medio de referencias personales.

2. **¿En cuanto tiempo aproximadamente se realiza el proceso?**

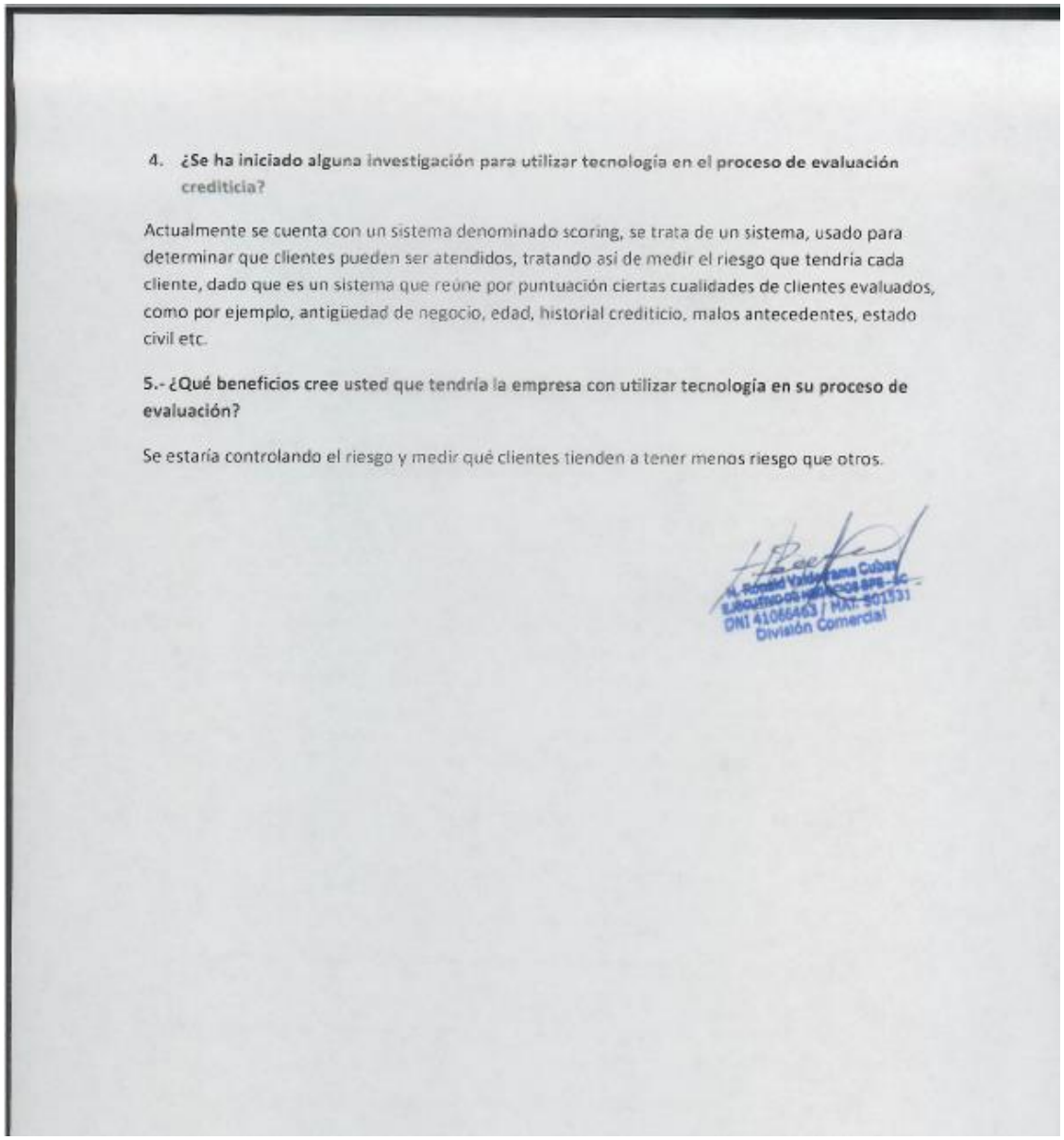
Para realizar los procesos de evaluación máximo 02 días.

3. **¿Qué medidas se realizan cuando un cliente no cumple con el pago de su préstamo?**

En primera instancia es trabajo del propio funcionario al quien es destinado dicho cliente para mantener un contacto de alerta ante cualquier situación de incumplimiento de pago: para lo cual se le brinda facilidades al cliente de :

- Reprogramar (siempre y cuando el cliente esté al dia e informe de que ha futuro tendría complicaciones en el cumplimiento. Se procede a una ampliación en el número de cuotas con la finalidad de que el monto de cuota baje y cliente tenga la mayor facilidad para cancelar, cabe indicar que la reprogramación es interna sin informar a las centrales de riesgo por lo que cliente no se veria mal informado en el sistema.
- Refinanciación (cliente con meses de inclumplimiento, en este caso el beneficio es el mismo del paso anterior con la diferencia que será informado en llas centrales de riego.)

Figura Nº 39 Entrevista
Fuente: Elaboración propia



4. **¿Se ha iniciado alguna investigación para utilizar tecnología en el proceso de evaluación crediticia?**

Actualmente se cuenta con un sistema denominado scoring, se trata de un sistema, usado para determinar que clientes pueden ser atendidos, tratando así de medir el riesgo que tendría cada cliente, dado que es un sistema que reúne por puntuación ciertas cualidades de clientes evaluados, como por ejemplo, antigüedad de negocio, edad, historial crediticio, malos antecedentes, estado civil etc.

5.- ¿Qué beneficios cree usted que tendría la empresa con utilizar tecnología en su proceso de evaluación?

Se estaría controlando el riesgo y medir qué clientes tienden a tener menos riesgo que otros.

Figura Nº 40 Entrevista
Fuente: Elaboración propia

ANEXO 02 – RESULTADOS DE PRUEBAS REALIZADO AL MODELO DE MACHINE LEARNING

Instancia	Valor actual	Valor de predicción	Error de predicción
1	2:yes	2:yes	0.956
2	2:yes	2:yes	0.506
3	1:no	2:yes (+)	0.925
4	2:yes	2:yes	0.981
5	2:yes	2:yes	0.965
6	2:yes	2:yes	0.971
7	2:yes	2:yes	0.961
8	2:yes	2:yes	0.972
9	2:yes	2:yes	0.940
10	2:yes	2:yes	0.942
11	2:yes	2:yes	0.940
12	2:yes	2:yes	0.845
13	2:yes	2:yes	0.962
14	2:yes	2:yes	0.984
15	2:yes	2:yes	0.961
16	2:yes	2:yes	0.924
17	2:yes	2:yes	0.948
18	2:yes	2:yes	0.987
19	2:yes	2:yes	0.973
20	2:yes	2:yes	0.833
21	2:yes	2:yes	0.630
22	2:yes	2:yes	0.976
23	2:yes	2:yes	0.997
24	2:yes	2:yes	0.963
25	1:no	2:yes (+)	0.653
26	2:yes	2:yes	0.731
27	2:yes	2:yes	0.934
28	1:no	2:yes (+)	0.944
29	2:yes	2:yes	0.963
30	2:yes	2:yes	0.846
31	2:yes	2:yes	0.972
32	2:yes	2:yes	0.776
33	2:yes	2:yes	0.850

34	2:yes	2:yes	0.976
35	2:yes	2:yes	0.837
36	2:yes	2:yes	0.782
37	1:no	2:yes (+)	0.979
38	1:no	2:yes (+)	0.629
39	2:yes	2:yes	0.985
40	2:yes	2:yes	0.621
41	2:yes	2:yes	0.883
42	1:no	2:yes (+)	0.643
43	2:yes	2:yes	0.934
44	2:yes	2:yes	0.601
45	1:no	1:no	0.669
46	2:yes	2:yes	0.767
47	1:no	2:yes (+)	0.973
48	2:yes	2:yes	0.591
49	2:yes	2:yes	0.932
50	2:yes	2:yes	0.977
51	2:yes	2:yes	0.739
52	2:yes	2:yes	0.960
53	2:yes	2:yes	0.980
54	2:yes	2:yes	0.977
55	2:yes	2:yes	0.699
56	2:yes	2:yes	0.682
57	2:yes	2:yes	0.962
58	2:yes	2:yes	0.967
59	2:yes	2:yes	0.915
60	2:yes	2:yes	0.958
61	2:yes	2:yes	0.817
62	2:yes	2:yes	0.767
63	2:yes	2:yes	0.929
64	2:yes	2:yes	0.724
65	2:yes	2:yes	0.939
66	2:yes	2:yes	0.984
67	2:yes	2:yes	0.666
68	2:yes	2:yes	0.969
69	2:yes	2:yes	0.914
70	2:yes	2:yes	0.780

71	2:yes	1:no (+)	0.655
72	2:yes	2:yes	0.806
73	2:yes	2:yes	0.709
74	1:no	2:yes (+)	0.825
75	2:yes	2:yes	0.818
76	2:yes	2:yes	0.701

ANEXO 3 – RESULTADOS DE LA APLICACIÓN INFORMÁTICA

Instancia	Valor actual	Valor de predicción
1	yes	yes
2	yes	yes
3	yes	yes
4	yes	yes
5	yes	yes
6	yes	yes
7	yes	yes
8	yes	yes
9	yes	yes
10	yes	yes
11	yes	yes
12	yes	yes
13	yes	yes
14	yes	yes
15	yes	yes
16	yes	yes
17	yes	yes
18	yes	yes
19	yes	yes
20	yes	yes
21	yes	yes
22	yes	yes
23	yes	yes
24	yes	yes
25	yes	yes
26	yes	yes
27	yes	yes
28	yes	yes
29	yes	yes
30	yes	yes
31	yes	yes
32	yes	yes
33	yes	yes
34	yes	yes
35	yes	yes
36	yes	yes
37	yes	yes
38	yes	no
39	yes	yes
40	yes	yes
41	yes	yes

42	yes	yes
43	yes	yes
44	yes	yes
45	yes	yes
46	yes	yes
47	yes	yes
48	yes	yes
49	yes	yes
50	yes	yes
51	yes	yes
52	yes	yes
53	yes	yes
54	yes	yes
55	yes	yes
56	yes	yes
57	yes	yes
58	yes	no
59	yes	yes
60	yes	yes
61	yes	yes
62	yes	yes
63	yes	yes
64	yes	yes
65	yes	yes
66	yes	yes
67	yes	yes
68	yes	yes
69	yes	yes
70	yes	yes
71	yes	yes
72	yes	yes
73	yes	yes
74	yes	yes
75	yes	yes
76	yes	yes
77	yes	yes
78	yes	yes
79	yes	yes
80	yes	yes
81	yes	yes
82	yes	yes
83	yes	yes
84	yes	yes

85	yes	yes
86	yes	yes
87	yes	yes
88	yes	yes
89	yes	yes
90	yes	yes
91	yes	yes
92	yes	yes
93	yes	yes
94	yes	yes
95	yes	yes
96	yes	yes
97	yes	yes
98	yes	yes
99	yes	yes
100	yes	yes
101	yes	yes
102	yes	yes
103	yes	yes
104	yes	yes
105	yes	yes
106	yes	yes
107	yes	yes
108	yes	yes
109	yes	yes
110	yes	yes
111	yes	yes
112	yes	yes
113	yes	yes
114	yes	yes
115	yes	yes
116	yes	yes
117	yes	yes
118	yes	yes
119	yes	yes
120	yes	yes
121	yes	yes
122	yes	yes
123	yes	yes
124	yes	yes
125	yes	yes
126	yes	yes
127	yes	yes

128	yes	yes
129	yes	yes
130	yes	yes
131	yes	yes
132	yes	yes
133	yes	yes
134	yes	yes
135	yes	yes
136	yes	yes
137	yes	yes
138	yes	yes
139	yes	yes
140	yes	yes
141	yes	yes
142	yes	yes
143	yes	yes
144	yes	yes
145	yes	yes
146	yes	yes
147	yes	yes
148	yes	yes
149	yes	yes
150	yes	yes
151	yes	yes
152	yes	yes
153	yes	yes
154	yes	yes
155	yes	yes
156	yes	yes
157	yes	yes
158	yes	yes
159	no	yes
160	no	yes
161	no	yes
162	no	yes
163	no	yes
164	no	yes
165	no	yes
166	no	no
167	no	no
168	no	yes
169	no	no
170	no	no

171	no	no
172	no	yes
173	no	no
174	no	yes
175	no	no
176	no	yes
177	no	yes
178	no	yes
179	no	yes
180	no	yes
181	no	no
182	no	no
183	no	yes
184	no	yes
185	no	yes
186	no	yes
187	no	yes
188	no	yes
189	no	yes
190	no	yes
191	no	yes
192	no	yes
193	no	no
194	no	no
195	no	no
196	no	yes
197	no	yes
198	no	yes
199	no	yes
200	no	yes
201	no	yes
202	no	yes
203	no	yes
204	no	yes
205	no	yes
206	no	yes
207	yes	yes
208	yes	yes
209	yes	yes
210	yes	yes
211	yes	yes
212	yes	yes
213	yes	yes

214	yes	yes
215	yes	yes
216	yes	yes
217	yes	yes
218	yes	yes
219	yes	yes
220	yes	yes
221	yes	yes
222	yes	yes
223	yes	yes
224	yes	yes
225	yes	yes
226	yes	yes
227	yes	yes
228	yes	yes
229	yes	yes
230	yes	yes
231	yes	yes
232	yes	yes
233	yes	yes
234	yes	yes
235	yes	yes
236	yes	yes
237	yes	yes
238	yes	yes
239	yes	yes
240	yes	yes
241	yes	yes
242	yes	yes
243	yes	yes
244	yes	yes
245	yes	yes
246	yes	yes
247	yes	yes
248	yes	yes
249	yes	yes
250	yes	no
251	yes	yes
252	yes	yes
253	yes	yes
254	yes	yes
255	yes	yes
256	yes	yes

257	yes	yes
258	yes	yes
259	yes	yes
260	yes	yes
261	yes	yes
262	yes	yes
263	yes	yes
264	yes	yes
265	yes	yes
266	yes	yes
267	yes	yes
268	yes	yes
269	yes	yes
270	yes	yes
271	yes	yes
272	yes	yes
273	yes	yes
274	yes	yes
275	yes	yes
276	yes	yes
277	yes	yes
278	yes	yes
279	yes	yes
280	yes	yes
281	yes	yes
282	yes	yes
283	yes	yes
284	yes	yes
285	yes	yes
286	yes	yes
287	yes	yes
288	yes	yes
289	yes	yes
290	yes	yes
291	yes	yes
292	yes	yes
293	yes	yes
294	yes	yes
295	yes	yes
296	yes	yes
297	yes	yes
298	yes	yes
299	yes	yes

300	yes	yes
301	yes	yes
302	yes	yes
303	yes	yes
304	yes	yes

ANEXO 04 – RESULTADOS DE LOS MONTOS GANADOS

Instancia	Valor Actual	Rentabilidad Actual	Valor de Predicción	Rentabilidad de Predicción
1	yes	2.457	yes	2.457
2	yes	5.068	yes	5.068
3	yes	2.017	yes	2.017
4	yes	2.231	yes	2.231
5	yes	38.888	yes	38.888
6	yes	10.614	yes	10.614
7	yes	10.894	yes	10.894
8	yes	10.794	yes	10.794
9	yes	32.742	yes	32.742
10	yes	1.743	yes	1.743
11	yes	41.531	yes	41.531
12	yes	2.953	yes	2.953
13	yes	52.744	yes	52.744
14	yes	1.673	yes	1.673
15	yes	14.789	yes	14.789
16	yes	4.019	yes	4.019
17	yes	3.026	yes	3.026
18	yes	5.959	yes	5.959
19	yes	4.133	yes	4.133
20	yes	18.697	yes	18.697
21	yes	2.336	yes	2.336
22	yes	0.417	yes	0.417
23	yes	7.448	yes	7.448
24	yes	3.026	yes	3.026
25	yes	3.188	yes	3.188
26	yes	3.188	yes	3.188
27	yes	6.38	yes	6.38
28	yes	54.619	yes	54.619
29	yes	7.795	yes	7.795
30	yes	23.363	yes	23.363
31	yes	7.69	yes	7.69
32	yes	0.496	yes	0.496
33	yes	8.267	yes	8.267
34	yes	92.779	yes	92.779
35	yes	20.95	yes	20.95
36	yes	18.217	yes	18.217
37	yes	2.977	yes	2.977
38	yes	82.933	no	0.0
39	yes	33.095	yes	33.095
40	yes	34.609	yes	34.609

41	yes	0.942	yes	0.942
42	yes	0.942	yes	0.942
43	yes	20.698	yes	20.698
44	yes	6.45	yes	6.45
45	yes	4.961	yes	4.961
46	yes	41.228	yes	41.228
47	yes	32.958	yes	32.958
48	yes	2.643	yes	2.643
49	yes	1.983	yes	1.983
50	yes	1.726	yes	1.726
51	yes	7.45	yes	7.45
52	yes	43.147	yes	43.147
53	yes	2.481	yes	2.481
54	yes	18.59	yes	18.59
55	yes	19.092	yes	19.092
56	yes	6.457	yes	6.457
57	yes	23.456	yes	23.456
58	yes	32.397	no	0.0
59	yes	10.54	yes	10.54
60	yes	21.604	yes	21.604
61	yes	8.874	yes	8.874
62	yes	4.465	yes	4.465
63	yes	48.096	yes	48.096
64	yes	25.157	yes	25.157
65	yes	6.912	yes	6.912
66	yes	14.872	yes	14.872
67	yes	21.604	yes	21.604
68	yes	3.927	yes	3.927
69	yes	69.19	yes	69.19
70	yes	25.157	yes	25.157
71	yes	14.808	yes	14.808
72	yes	58.303	yes	58.303
73	yes	21.604	yes	21.604
74	yes	24.69	yes	24.69
75	yes	21.604	yes	21.604
76	yes	29.769	yes	29.769
77	yes	2.977	yes	2.977
78	yes	40.839	yes	40.839
79	yes	21.604	yes	21.604
80	yes	80.663	yes	80.663
81	yes	4.691	yes	4.691
82	yes	45.238	yes	45.238
83	yes	43.147	yes	43.147

84	yes	29.641	yes	29.641
85	yes	6.32	yes	6.32
86	yes	21.604	yes	21.604
87	yes	4.614	yes	4.614
88	yes	32.883	yes	32.883
89	yes	44.461	yes	44.461
90	yes	21.604	yes	21.604
91	yes	21.604	yes	21.604
92	yes	26.476	yes	26.476
93	yes	44.461	yes	44.461
94	yes	32.686	yes	32.686
95	yes	2.874	yes	2.874
96	yes	12.251	yes	12.251
97	yes	20.695	yes	20.695
98	yes	13.786	yes	13.786
99	yes	7.188	yes	7.188
100	yes	24.662	yes	24.662
101	yes	8.24	yes	8.24
102	yes	2.479	yes	2.479
103	yes	10.3	yes	10.3
104	yes	12.36	yes	12.36
105	yes	53.38	yes	53.38
106	yes	6.555	yes	6.555
107	yes	12.944	yes	12.944
108	yes	6.18	yes	6.18
109	yes	10.986	yes	10.986
110	yes	9.2	yes	9.2
111	yes	6.18	yes	6.18
112	yes	9.2	yes	9.2
113	yes	10.465	yes	10.465
114	yes	12.364	yes	12.364
115	yes	6.974	yes	6.974
116	yes	6.32	yes	6.32
117	yes	69.19	yes	69.19
118	yes	2.982	yes	2.982
119	yes	27.712	yes	27.712
120	yes	45.238	yes	45.238
121	yes	7.795	yes	7.795
122	yes	31.259	yes	31.259
123	yes	16.441	yes	16.441
124	yes	3.141	yes	3.141
125	yes	3.141	yes	3.141
126	yes	2.977	yes	2.977

127	yes	7.943	yes	7.943
128	yes	9.746	yes	9.746
129	yes	7.188	yes	7.188
130	yes	41.391	yes	41.391
131	yes	4.699	yes	4.699
132	yes	7.442	yes	7.442
133	yes	6.555	yes	6.555
134	yes	1.85	yes	1.85
135	yes	34.471	yes	34.471
136	yes	39.646	yes	39.646
137	yes	61.852	yes	61.852
138	yes	43.297	yes	43.297
139	yes	1.7	yes	1.7
140	yes	37.111	yes	37.111
141	yes	8.278	yes	8.278
142	yes	3.634	yes	3.634
143	yes	2.478	yes	2.478
144	yes	18.733	yes	18.733
145	yes	13.153	yes	13.153
146	yes	3.605	yes	3.605
147	yes	15.858	yes	15.858
148	yes	4.313	yes	4.313
149	yes	9.981	yes	9.981
150	yes	39.931	yes	39.931
151	yes	3.594	yes	3.594
152	yes	30.271	yes	30.271
153	yes	13.223	yes	13.223
154	yes	7.436	yes	7.436
155	yes	103.702	yes	103.702
156	yes	7.533	yes	7.533
157	yes	2.53	yes	2.53
158	yes	5.414	yes	5.414
159	no	0.0	yes	5.022
160	no	0.0	yes	8.26
161	no	0.0	yes	6.127
162	no	0.0	yes	5.022
163	no	0.0	yes	20.099
164	no	0.0	yes	20.636
165	no	0.0	yes	32.771
166	no	0.0	no	0.0
167	no	0.0	no	0.0
168	no	0.0	yes	27.213
169	no	0.0	no	0.0

170	no	0.0	no	0.0
171	no	0.0	no	0.0
172	no	0.0	yes	0
173	no	0.0	no	0.0
174	no	0.0	yes	0
175	no	0.0	no	0.0
176	no	0.0	yes	12.639
177	no	0.0	yes	20.039
178	no	0.0	yes	8.267
179	no	0.0	yes	8.26
180	no	0.0	yes	53.852
181	no	0.0	no	0.0
182	no	0.0	no	0.0
183	no	0.0	yes	7.851
184	no	0.0	yes	3.924
185	no	0.0	yes	9.3
186	no	0.0	yes	16.215
187	no	0.0	yes	112.71
188	no	0.0	yes	25.946
189	no	0.0	yes	22.568
190	no	0.0	yes	3.304
191	no	0.0	yes	3.304
192	no	0.0	yes	120.76
193	no	0.0	no	0.0
194	no	0.0	no	0.0
195	no	0.0	no	0.0
196	no	0.0	yes	36.394
197	no	0.0	yes	19.744
198	no	0.0	yes	16.328
199	no	0.0	yes	55.309
200	no	0.0	yes	5.462
201	no	0.0	yes	40.7
202	no	0.0	yes	2.895
203	no	0.0	yes	50.08
204	no	0.0	yes	21.667
205	no	0.0	yes	35.311
206	no	0.0	yes	20.099
207	yes	32.771	yes	32.771
208	yes	11.028	yes	11.028
209	yes	21.756	yes	21.756
210	yes	8.685	yes	8.685
211	yes	13.153	yes	13.153
212	yes	5.061	yes	5.061

213	yes	32.263	yes	32.263
214	yes	2.727	yes	2.727
215	yes	33.008	yes	33.008
216	yes	5.373	yes	5.373
217	yes	2.727	yes	2.727
218	yes	2.875	yes	2.875
219	yes	0.991	yes	0.991
220	yes	20.53	yes	20.53
221	yes	2.974	yes	2.974
222	yes	1.983	yes	1.983
223	yes	6.569	yes	6.569
224	yes	8.934	yes	8.934
225	yes	15.109	yes	15.109
226	yes	8.685	yes	8.685
227	yes	7.188	yes	7.188
228	yes	12.345	yes	12.345
229	yes	6.197	yes	6.197
230	yes	8.434	yes	8.434
231	yes	65.93	yes	65.93
232	yes	17.131	yes	17.131
233	yes	65.93	yes	65.93
234	yes	46.889	yes	46.889
235	yes	17.876	yes	17.876
236	yes	72.298	yes	72.298
237	yes	11.094	yes	11.094
238	yes	6.2	yes	6.2
239	yes	10.651	yes	10.651
240	yes	10.339	yes	10.339
241	yes	23.384	yes	23.384
242	yes	23.16	yes	23.16
243	yes	3.927	yes	3.927
244	yes	7.442	yes	7.442
245	yes	11.168	yes	11.168
246	yes	36.872	yes	36.872
247	yes	24.807	yes	24.807
248	yes	8.629	yes	8.629
249	yes	16.52	yes	16.52
250	yes	43.425	no	0.0
251	yes	2.478	yes	2.478
252	yes	5.414	yes	5.414
253	yes	9.2	yes	9.2
254	yes	10.834	yes	10.834
255	yes	63.084	yes	63.084

256	yes	33.152	yes	33.152
257	yes	4.298	yes	4.298
258	yes	29.239	yes	29.239
259	yes	18.993	yes	18.993
260	yes	21.228	yes	21.228
261	yes	20.039	yes	20.039
262	yes	7.442	yes	7.442
263	yes	57.585	yes	57.585
264	yes	19.037	yes	19.037
265	yes	30.116	yes	30.116
266	yes	20.844	yes	20.844
267	yes	29.123	yes	29.123
268	yes	2.314	yes	2.314
269	yes	8.874	yes	8.874
270	yes	1.652	yes	1.652
271	yes	4.961	yes	4.961
272	yes	1.653	yes	1.653
273	yes	1.686	yes	1.686
274	yes	23.384	yes	23.384
275	yes	4.961	yes	4.961
276	yes	16.268	yes	16.268
277	yes	5.823	yes	5.823
278	yes	0.744	yes	0.744
279	yes	1.983	yes	1.983
280	yes	26.303	yes	26.303
281	yes	17.876	yes	17.876
282	yes	43.23	yes	43.23
283	yes	3.306	yes	3.306
284	yes	43.297	yes	43.297
285	yes	8.638	yes	8.638
286	yes	2.953	yes	2.953
287	yes	9.291	yes	9.291
288	yes	4.133	yes	4.133
289	yes	8.685	yes	8.685
290	yes	16.014	yes	16.014
291	yes	3.307	yes	3.307
292	yes	5.745	yes	5.745
293	yes	17.235	yes	17.235
294	yes	12.298	yes	12.298
295	yes	3.139	yes	3.139
296	yes	10.171	yes	10.171
297	yes	37.051	yes	37.051
298	yes	30.253	yes	30.253

299	yes	3.927	yes	3.927
300	yes	18.08	yes	18.08
301	yes	4.711	yes	4.711
302	yes	6.893	yes	6.893
303	yes	11.831	yes	11.831
304	yes	37.472	yes	37.472

ANEXO 05 – RESULTADOS DE INDICADOR DE RIESGOS

Instancia	Valor Actual	Riesgo Actual	Valor predicción	Riesgo de Predicción
1	yes	25.0	yes	25.0
2	yes	76.65	yes	76.65
3	yes	30.5	yes	30.5
4	yes	33.75	yes	33.75
5	yes	150.0	yes	150.0
6	yes	71.25	yes	71.25
7	yes	59.8	yes	59.8
8	yes	59.25	yes	59.25
9	yes	152.0	yes	152.0
10	yes	35.15	yes	35.15
11	yes	210.0	yes	210.0
12	yes	25.5	yes	25.5
13	yes	200.0	yes	200.0
14	yes	25.3	yes	25.3
15	yes	50.0	yes	50.0
16	yes	40.5	yes	40.5
17	yes	30.5	yes	30.5
18	yes	40.0	yes	40.0
19	yes	50.0	yes	50.0
20	yes	150.0	yes	150.0
21	yes	25.0	yes	25.0
22	yes	12.63	yes	12.63
23	yes	50.0	yes	50.0
24	yes	30.5	yes	30.5
25	yes	96.5	yes	96.5
26	yes	96.5	yes	96.5
27	yes	96.5	yes	96.5
28	yes	250.0	yes	250.0
29	yes	50.0	yes	50.0
30	yes	128.25	yes	128.25
31	yes	77.5	yes	77.5
32	yes	15.0	yes	15.0
33	yes	100.0	yes	100.0
34	yes	375.0	yes	375.0
35	yes	115.0	yes	115.0
36	yes	100.0	yes	100.0
37	yes	30.0	yes	30.0
38	yes	385.0	no	0.0
39	yes	250.0	yes	250.0
40	yes	175.0	yes	175.0

41	yes	19.0	yes	19.0
42	yes	19.0	yes	19.0
43	yes	60.0	yes	60.0
44	yes	65.0	yes	65.0
45	yes	50.0	yes	50.0
46	yes	500.0	yes	500.0
47	yes	153.0	yes	153.0
48	yes	80.0	yes	80.0
49	yes	40.0	yes	40.0
50	yes	26.1	yes	26.1
51	yes	45.0	yes	45.0
52	yes	250.0	yes	250.0
53	yes	25.0	yes	25.0
54	yes	375.0	yes	375.0
55	yes	42.5	yes	42.5
56	yes	39.0	yes	39.0
57	yes	190.0	yes	190.0
58	yes	490.0	no	0.0
59	yes	127.5	yes	127.5
60	yes	175.0	yes	175.0
61	yes	30.0	yes	30.0
62	yes	45.0	yes	45.0
63	yes	100.0	yes	100.0
64	yes	175.0	yes	175.0
65	yes	150.0	yes	150.0
66	yes	300.0	yes	300.0
67	yes	175.0	yes	175.0
68	yes	47.5	yes	47.5
69	yes	400.0	yes	400.0
70	yes	175.0	yes	175.0
71	yes	150.0	yes	150.0
72	yes	150.0	yes	150.0
73	yes	175.0	yes	175.0
74	yes	200.0	yes	200.0
75	yes	175.0	yes	175.0
76	yes	300.0	yes	300.0
77	yes	30.0	yes	30.0
78	yes	150.0	yes	150.0
79	yes	175.0	yes	175.0
80	yes	340.0	yes	340.0
81	yes	70.95	yes	70.95
82	yes	125.0	yes	125.0
83	yes	250.0	yes	250.0

84	yes	200.0	yes	200.0
85	yes	27.5	yes	27.5
86	yes	175.0	yes	175.0
87	yes	46.5	yes	46.5
88	yes	100.0	yes	100.0
89	yes	300.0	yes	300.0
90	yes	175.0	yes	175.0
91	yes	175.0	yes	175.0
92	yes	200.0	yes	200.0
93	yes	300.0	yes	300.0
94	yes	101.5	yes	101.5
95	yes	25.0	yes	25.0
96	yes	100.0	yes	100.0
97	yes	125.0	yes	125.0
98	yes	100.0	yes	100.0
99	yes	50.0	yes	50.0
100	yes	75.0	yes	75.0
101	yes	200.0	yes	200.0
102	yes	37.5	yes	37.5
103	yes	250.0	yes	250.0
104	yes	300.0	yes	300.0
105	yes	225.0	yes	225.0
106	yes	50.0	yes	50.0
107	yes	75.0	yes	75.0
108	yes	150.0	yes	150.0
109	yes	51.0	yes	51.0
110	yes	50.5	yes	50.5
111	yes	150.0	yes	150.0
112	yes	50.5	yes	50.5
113	yes	52.5	yes	52.5
114	yes	200.0	yes	200.0
115	yes	151.35	yes	151.35
116	yes	27.5	yes	27.5
117	yes	400.0	yes	400.0
118	yes	25.75	yes	25.75
119	yes	150.0	yes	150.0
120	yes	125.0	yes	125.0
121	yes	50.0	yes	50.0
122	yes	100.0	yes	100.0
123	yes	50.0	yes	50.0
124	yes	47.5	yes	47.5
125	yes	47.5	yes	47.5
126	yes	30.0	yes	30.0

127	yes	60.0	yes	60.0
128	yes	53.5	yes	53.5
129	yes	50.0	yes	50.0
130	yes	250.0	yes	250.0
131	yes	35.5	yes	35.5
132	yes	75.0	yes	75.0
133	yes	50.0	yes	50.0
134	yes	25.0	yes	25.0
135	yes	150.0	yes	150.0
136	yes	125.0	yes	125.0
137	yes	250.0	yes	250.0
138	yes	175.0	yes	175.0
139	yes	27.5	yes	27.5
140	yes	150.0	yes	150.0
141	yes	50.0	yes	50.0
142	yes	110.0	yes	110.0
143	yes	75.0	yes	75.0
144	yes	65.0	yes	65.0
145	yes	40.0	yes	40.0
146	yes	27.5	yes	27.5
147	yes	50.0	yes	50.0
148	yes	30.0	yes	30.0
149	yes	67.0	yes	67.0
150	yes	135.0	yes	135.0
151	yes	25.0	yes	25.0
152	yes	175.0	yes	175.0
153	yes	200.0	yes	200.0
154	yes	150.0	yes	150.0
155	yes	400.0	yes	400.0
156	yes	45.5	yes	45.5
157	yes	25.5	yes	25.5
158	yes	27.5	yes	27.5
159	no	0.0	yes	0.0
160	no	0.0	yes	0.0
161	no	0.0	yes	0.0
162	no	0.0	yes	0.0
163	no	0.0	yes	0.0
164	no	0.0	yes	0.0
165	no	0.0	yes	0.0
166	no	0.0	no	0.0
167	no	0.0	no	0.0
168	no	0.0	yes	0.0
169	no	0.0	no	0.0

170	no	0.0	no	0.0
171	no	0.0	no	0.0
172	no	0.0	yes	0.0
173	no	0.0	no	0.0
174	no	0.0	yes	0.0
175	no	0.0	no	0.0
176	no	0.0	yes	0.0
177	no	0.0	yes	0.0
178	no	0.0	yes	0.0
179	no	0.0	yes	0.0
180	no	0.0	yes	0.0
181	no	0.0	no	0.0
182	no	0.0	no	0.0
183	no	0.0	yes	0.0
184	no	0.0	yes	0.0
185	no	0.0	yes	0.0
186	no	0.0	yes	0.0
187	no	0.0	yes	0.0
188	no	0.0	yes	0.0
189	no	0.0	yes	0.0
190	no	0.0	yes	0.0
191	no	0.0	yes	0.0
192	no	0.0	yes	0.0
193	no	0.0	no	0.0
194	no	0.0	no	0.0
195	no	0.0	no	0.0
196	no	0.0	yes	0.0
197	no	0.0	yes	0.0
198	no	0.0	yes	0.0
199	no	0.0	yes	0.0
200	no	0.0	yes	0.0
201	no	0.0	yes	0.0
202	no	0.0	yes	0.0
203	no	0.0	yes	0.0
204	no	0.0	yes	0.0
205	no	0.0	yes	0.0
206	no	0.0	yes	0.0
207	yes	150.0	yes	150.0
208	yes	80.0	yes	80.0
209	yes	101.0	yes	101.0
210	yes	75.0	yes	75.0
211	yes	40.0	yes	40.0
212	yes	51.0	yes	51.0

213	yes	175.0	yes	175.0
214	yes	55.0	yes	55.0
215	yes	102.5	yes	102.5
216	yes	65.0	yes	65.0
217	yes	55.0	yes	55.0
218	yes	58.0	yes	58.0
219	yes	30.0	yes	30.0
220	yes	206.9	yes	206.9
221	yes	60.0	yes	60.0
222	yes	40.0	yes	40.0
223	yes	25.0	yes	25.0
224	yes	34.0	yes	34.0
225	yes	75.0	yes	75.0
226	yes	75.0	yes	75.0
227	yes	50.0	yes	50.0
228	yes	100.0	yes	100.0
229	yes	125.0	yes	125.0
230	yes	85.0	yes	85.0
231	yes	250.0	yes	250.0
232	yes	115.0	yes	115.0
233	yes	250.0	yes	250.0
234	yes	150.0	yes	150.0
235	yes	120.0	yes	120.0
236	yes	350.0	yes	350.0
237	yes	51.5	yes	51.5
238	yes	75.0	yes	75.0
239	yes	64.5	yes	64.5
240	yes	75.0	yes	75.0
241	yes	150.0	yes	150.0
242	yes	200.0	yes	200.0
243	yes	47.5	yes	47.5
244	yes	75.0	yes	75.0
245	yes	42.5	yes	42.5
246	yes	200.0	yes	200.0
247	yes	250.0	yes	250.0
248	yes	50.0	yes	50.0
249	yes	500.0	yes	500.0
250	yes	375.0	no	0.0
251	yes	75.0	yes	75.0
252	yes	27.5	yes	27.5
253	yes	50.5	yes	50.5
254	yes	75.0	yes	75.0
255	yes	155.0	yes	155.0

256	yes	153.9	yes	153.9
257	yes	65.0	yes	65.0
258	yes	252.5	yes	252.5
259	yes	127.5	yes	127.5
260	yes	142.5	yes	142.5
261	yes	110.0	yes	110.0
262	yes	75.0	yes	75.0
263	yes	250.0	yes	250.0
264	yes	104.5	yes	104.5
265	yes	227.5	yes	227.5
266	yes	180.0	yes	180.0
267	yes	220.0	yes	220.0
268	yes	35.0	yes	35.0
269	yes	30.0	yes	30.0
270	yes	50.0	yes	50.0
271	yes	50.0	yes	50.0
272	yes	25.0	yes	25.0
273	yes	25.5	yes	25.5
274	yes	150.0	yes	150.0
275	yes	50.0	yes	50.0
276	yes	55.0	yes	55.0
277	yes	176.25	yes	176.25
278	yes	15.0	yes	15.0
279	yes	30.0	yes	30.0
280	yes	250.0	yes	250.0
281	yes	120.0	yes	120.0
282	yes	150.0	yes	150.0
283	yes	50.0	yes	50.0
284	yes	175.0	yes	175.0
285	yes	37.5	yes	37.5
286	yes	25.5	yes	25.5
287	yes	51.0	yes	51.0
288	yes	50.0	yes	50.0
289	yes	75.0	yes	75.0
290	yes	107.5	yes	107.5
291	yes	40.0	yes	40.0
292	yes	25.0	yes	25.0
293	yes	75.0	yes	75.0
294	yes	75.0	yes	75.0
295	yes	95.0	yes	95.0
296	yes	102.5	yes	102.5
297	yes	250.0	yes	250.0
298	yes	261.25	yes	261.25

299	yes	47.5	yes	47.5
300	yes	75.0	yes	75.0
301	yes	71.25	yes	71.25
302	yes	50.0	yes	50.0
303	yes	40.0	yes	40.0
304	yes	150.0	yes	150.0

ANEXO 06 – RESULTADOS DE INDICADOR DE TIEMPO

Instancia	Valor actual	Tiempo actual (Segundos)	Valor de predicción	Tiempo de predicción (Segundos)
1	yes	172800	yes	1.0506
2	yes	259200	yes	1.1408
3	yes	86400	yes	0.6005
4	yes	259200	yes	0.5530
5	yes	86400	yes	0.3944
6	yes	86400	yes	1.0282
7	yes	259200	yes	0.7687
8	yes	259200	yes	0.6439
9	yes	259200	yes	0.4961
10	yes	259200	yes	0.8849
11	yes	259200	yes	0.1763
12	yes	86400	yes	0.9155
13	yes	172800	yes	0.9600
14	yes	86400	yes	0.8049
15	yes	86400	yes	0.3476
16	yes	259200	yes	0.3821
17	yes	172800	yes	1.1373
18	yes	86400	yes	1.1249
19	yes	259200	yes	0.5617
20	yes	172800	yes	0.6019
21	yes	259200	yes	0.5765
22	yes	86400	yes	0.2111
23	yes	259200	yes	0.4183
24	yes	86400	yes	1.0037
25	yes	86400	yes	0.9221
26	yes	172800	yes	1.0373
27	yes	86400	yes	0.4134
28	yes	86400	yes	0.2296
29	yes	172800	yes	0.4942
30	yes	172800	yes	0.1939
31	yes	86400	yes	0.5406
32	yes	259200	yes	1.0332
33	yes	259200	yes	0.8414
34	yes	259200	yes	0.2633
35	yes	172800	yes	1.0007
36	yes	172800	yes	0.2388
37	yes	259200	yes	0.2269
38	yes	172800	no	0.2209
39	yes	86400	yes	0.1586

40	yes	172800	yes	0.5250
41	yes	259200	yes	0.3825
42	yes	172800	yes	0.5996
43	yes	86400	yes	0.6130
44	yes	172800	yes	0.5556
45	yes	172800	yes	0.6450
46	yes	86400	yes	1.1312
47	yes	172800	yes	0.4993
48	yes	259200	yes	0.2510
49	yes	86400	yes	0.7945
50	yes	172800	yes	0.4323
51	yes	259200	yes	0.2059
52	yes	172800	yes	0.9350
53	yes	172800	yes	1.0814
54	yes	86400	yes	1.0076
55	yes	259200	yes	0.6234
56	yes	172800	yes	0.8857
57	yes	172800	yes	0.4681
58	yes	172800	no	0.6263
59	yes	172800	yes	0.5354
60	yes	259200	yes	0.2666
61	yes	259200	yes	0.8878
62	yes	172800	yes	0.2620
63	yes	259200	yes	0.1839
64	yes	172800	yes	0.7933
65	yes	172800	yes	0.2185
66	yes	259200	yes	0.7648
67	yes	172800	yes	0.4363
68	yes	86400	yes	0.5893
69	yes	259200	yes	0.4268
70	yes	259200	yes	1.0524
71	yes	86400	yes	0.9190
72	yes	172800	yes	1.1377
73	yes	172800	yes	0.4073
74	yes	86400	yes	0.2381
75	yes	172800	yes	0.8288
76	yes	172800	yes	0.2327
77	yes	259200	yes	0.9941
78	yes	259200	yes	0.6963
79	yes	86400	yes	1.1274
80	yes	172800	yes	0.4453
81	yes	86400	yes	0.2733
82	yes	86400	yes	0.8133

83	yes	259200	yes	1.0915
84	yes	172800	yes	0.5832
85	yes	172800	yes	0.7617
86	yes	86400	yes	0.7102
87	yes	172800	yes	0.2664
88	yes	259200	yes	1.0198
89	yes	172800	yes	0.1699
90	yes	259200	yes	1.1448
91	yes	172800	yes	0.4472
92	yes	86400	yes	1.1341
93	yes	259200	yes	0.7454
94	yes	86400	yes	0.8723
95	yes	86400	yes	0.5575
96	yes	86400	yes	0.5680
97	yes	259200	yes	0.9624
98	yes	172800	yes	0.7036
99	yes	172800	yes	0.5851
100	yes	86400	yes	1.0881
101	yes	172800	yes	0.3913
102	yes	86400	yes	1.0639
103	yes	172800	yes	0.7660
104	yes	172800	yes	0.1771
105	yes	259200	yes	0.3394
106	yes	259200	yes	0.7097
107	yes	259200	yes	0.4599
108	yes	172800	yes	1.0983
109	yes	86400	yes	0.1659
110	yes	259200	yes	0.3066
111	yes	172800	yes	1.0269
112	yes	259200	yes	0.7094
113	yes	86400	yes	0.8711
114	yes	86400	yes	0.7147
115	yes	172800	yes	0.6842
116	yes	259200	yes	0.4096
117	yes	259200	yes	1.1029
118	yes	172800	yes	0.9452
119	yes	259200	yes	0.3220
120	yes	172800	yes	0.1629
121	yes	259200	yes	1.1394
122	yes	259200	yes	0.3764
123	yes	172800	yes	0.2780
124	yes	86400	yes	0.7103
125	yes	259200	yes	1.0826

126	yes	172800	yes	0.7337
127	yes	86400	yes	0.9292
128	yes	259200	yes	0.1838
129	yes	259200	yes	0.3196
130	yes	259200	yes	0.6176
131	yes	172800	yes	1.0197
132	yes	86400	yes	0.6394
133	yes	259200	yes	0.8149
134	yes	86400	yes	0.5678
135	yes	259200	yes	0.4234
136	yes	259200	yes	0.7441
137	yes	86400	yes	0.4450
138	yes	259200	yes	0.6219
139	yes	172800	yes	0.2389
140	yes	86400	yes	0.8763
141	yes	86400	yes	0.5059
142	yes	86400	yes	0.9189
143	yes	172800	yes	0.7269
144	yes	172800	yes	0.7208
145	yes	86400	yes	0.9281
146	yes	86400	yes	0.8398
147	yes	86400	yes	0.3993
148	yes	172800	yes	0.8583
149	yes	259200	yes	0.7941
150	yes	172800	yes	0.3688
151	yes	86400	yes	0.1818
152	yes	86400	yes	0.4637
153	yes	259200	yes	0.8903
154	yes	86400	yes	0.4476
155	yes	259200	yes	0.8687
156	yes	86400	yes	1.1090
157	yes	172800	yes	0.2898
158	yes	86400	yes	1.0061
159	no	86400	yes	0.5317
160	no	259200	yes	0.9026
161	no	172800	yes	1.0529
162	no	259200	yes	0.6485
163	no	259200	yes	0.7666
164	no	86400	yes	1.0051
165	no	259200	yes	0.7559
166	no	86400	no	1.0513
167	no	86400	no	0.5712
168	no	86400	yes	0.7644

169	no	86400	no	0.3382
170	no	172800	no	0.9129
171	no	172800	no	0.6554
172	no	259200	yes	0.6510
173	no	259200	no	0.3451
174	no	259200	yes	0.7414
175	no	172800	no	0.7870
176	no	172800	yes	1.0414
177	no	259200	yes	0.5235
178	no	86400	yes	0.1617
179	no	259200	yes	0.4695
180	no	172800	yes	0.6325
181	no	259200	no	0.3542
182	no	172800	no	0.6214
183	no	86400	yes	1.1337
184	no	172800	yes	0.2177
185	no	259200	yes	0.9635
186	no	172800	yes	0.4061
187	no	259200	yes	0.1767
188	no	172800	yes	0.8243
189	no	259200	yes	0.5749
190	no	172800	yes	0.7000
191	no	259200	yes	0.2048
192	no	259200	yes	0.6244
193	no	259200	no	0.5560
194	no	86400	no	0.6443
195	no	172800	no	1.1256
196	no	86400	yes	1.1236
197	no	86400	yes	0.9845
198	no	172800	yes	0.5882
199	no	259200	yes	0.3800
200	no	259200	yes	0.6527
201	no	259200	yes	0.7982
202	no	259200	yes	0.7722
203	no	172800	yes	0.2825
204	no	259200	yes	1.1252
205	no	172800	yes	0.2035
206	no	172800	yes	0.8705
207	yes	86400	yes	0.1771
208	yes	86400	yes	0.5981
209	yes	259200	yes	1.0815
210	yes	172800	yes	0.5447
211	yes	86400	yes	0.8598

212	yes	86400	yes	0.3775
213	yes	172800	yes	0.5444
214	yes	259200	yes	1.0589
215	yes	86400	yes	0.7308
216	yes	86400	yes	1.0268
217	yes	172800	yes	1.0964
218	yes	86400	yes	0.3442
219	yes	172800	yes	0.6356
220	yes	259200	yes	0.7199
221	yes	259200	yes	0.2844
222	yes	86400	yes	0.2369
223	yes	86400	yes	1.0046
224	yes	259200	yes	0.5513
225	yes	259200	yes	0.8989
226	yes	172800	yes	0.9693
227	yes	259200	yes	1.1032
228	yes	259200	yes	0.4709
229	yes	86400	yes	1.1192
230	yes	86400	yes	0.2687
231	yes	172800	yes	0.9210
232	yes	172800	yes	0.1646
233	yes	172800	yes	1.0505
234	yes	86400	yes	0.7190
235	yes	259200	yes	1.1054
236	yes	172800	yes	0.5966
237	yes	259200	yes	0.6811
238	yes	86400	yes	0.9436
239	yes	86400	yes	0.7365
240	yes	86400	yes	1.0860
241	yes	259200	yes	0.4314
242	yes	259200	yes	0.3701
243	yes	86400	yes	0.4538
244	yes	86400	yes	0.3707
245	yes	86400	yes	0.9241
246	yes	86400	yes	0.2864
247	yes	172800	yes	0.9814
248	yes	259200	yes	0.4682
249	yes	86400	yes	0.4821
250	yes	259200	no	0.2917
251	yes	172800	yes	1.0316
252	yes	259200	yes	0.7383
253	yes	172800	yes	0.4294
254	yes	172800	yes	0.3301

255	yes	86400	yes	0.8236
256	yes	259200	yes	0.1541
257	yes	86400	yes	0.7661
258	yes	172800	yes	0.2564
259	yes	86400	yes	0.6349
260	yes	259200	yes	1.0687
261	yes	259200	yes	0.4433
262	yes	259200	yes	0.5356
263	yes	86400	yes	0.3856
264	yes	259200	yes	0.4173
265	yes	172800	yes	0.2259
266	yes	172800	yes	0.8654
267	yes	172800	yes	0.7744
268	yes	259200	yes	1.1271
269	yes	172800	yes	0.4224
270	yes	86400	yes	0.2391
271	yes	172800	yes	0.5107
272	yes	172800	yes	0.4216
273	yes	259200	yes	0.9561
274	yes	259200	yes	0.6942
275	yes	259200	yes	0.7347
276	yes	172800	yes	0.4238
277	yes	259200	yes	1.0214
278	yes	172800	yes	1.1029
279	yes	172800	yes	0.6279
280	yes	172800	yes	0.7752
281	yes	259200	yes	0.3688
282	yes	259200	yes	0.1587
283	yes	172800	yes	1.0898
284	yes	86400	yes	1.0824
285	yes	86400	yes	0.8952
286	yes	259200	yes	0.4091
287	yes	172800	yes	1.0928
288	yes	172800	yes	0.8481
289	yes	259200	yes	0.6032
290	yes	86400	yes	0.3576
291	yes	86400	yes	0.9896
292	yes	172800	yes	0.3278
293	yes	259200	yes	0.6317
294	yes	172800	yes	0.3902
295	yes	86400	yes	0.8048
296	yes	259200	yes	0.1641
297	yes	172800	yes	0.9309

298	yes	259200	yes	1.1398
299	yes	259200	yes	0.7225
300	yes	259200	yes	0.4257
301	yes	172800	yes	0.2713
302	yes	259200	yes	1.0560
303	yes	86400	yes	0.1807