Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería Mecánica Eléctrica

DISEÑO DE INVESTIGACIÓN DE UNA HERRAMIENTA PARA LA ESTIMACIÓN DE LA PROBABILIDAD DE ÉXITO DE UN NEGOCIO EN UNA EMPRESA DE DISTRIBUCIÓN DE PRODUCTOS UTILIZANDO MÉTODOS DE APRENDIZAJE AUTOMÁTICO

Juan Carlos Martini Palma

Asesorado por Inga. Lilia Susana Beltrán Paiz

Guatemala, julio de 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

DISEÑO DE INVESTIGACIÓN DE UNA HERRAMIENTA PARA LA ESTIMACIÓN DE LA PROBABILIDAD DE ÉXITO DE UN NEGOCIO EN UNA EMPRESA DE DISTRIBUCIÓN DE PRODUCTOS UTILIZANDO MÉTODOS DE APRENDIZAJE AUTOMÁTICO

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

JUAN CARLOS MARTINI PALMA
ASESORADO POR INGA. LILIA SUSANA BELTRÁN PAIZ

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN ELECTRÓNICA

GUATEMALA, JULIO DE 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Inga. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Vladimir Armando Cruz Lorente
VOCAL V	Br. Fernando José de Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANA	Ing. Aurelia Anabela Córdova Estrada
EXAMINADOR	Ing. Guillermo Antonio Puente Romero
EXAMINADOR	Ing. Luis Eduardo Durán Córdova
EXAMINADOR	Ing. Marvin Marino Hernández Fernández
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

DISEÑO DE INVESTIGACIÓN DE UNA HERRAMIENTA PARA LA ESTIMACIÓN DE LA PROBABILIDAD DE ÉXITO DE UN NEGOCIO EN UNA EMPRESA DE DISTRIBUCIÓN DE PRODUCTOS UTILIZANDO MÉTODOS DE APRENDIZAJE AUTOMÁTICO

Tema que me fuera asignado por la Dirección de Escuela de Estudios de Postgrado con fecha 19 de noviembre de 2021.

Juan Carlos Martini Palma

EEPFI-PP-0147-2022
Guatemala, 12 de enero de 2022

Director
Armando Alonso Rivera Carrillo
Escuela De Ingenieria Mecanica Electrica
Presente.

Estimado Ing. Rivera

Reciba un cordial saludo de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería.

El propósito de la presente es para informarle que se ha revisado y aprobado el Diseño de Investigación titulado: **DISEÑO DE UNA RED NEURONAL PARA LA ESTIMACIÓN DE LA PROBABILIDAD DE ÉXITO DE UN NEGOCIO EN UNA EMPRESA DE DISTRIBUCIÓN DE PRODUCTOS.**, el cual se enmarca en la línea de investigación: **Análisis de datos - Análisis de datos**, presentado por el estudiante **Juan Carlos Martini Palma** carné número **201408417**, quien optó por la modalidad del "PROCESO DE GRADUACIÓN DE LOS ESTUDIANTES DE LA FACULTAD DE INGENIERÍA OPCIÓN ESTUDIOS DE POSTGRADO". Previo a culminar sus estudios en la Maestría en ARTES en Ingeniería Para La Industria Con Especialidad En Ciencias De La Computación.

Y habiendo cumplido y aprobado con los requisitos establecidos en el normativo de este Proceso de Graduación en el Punto 6.2, aprobado por la Junta Directiva de la Facultad de Ingeniería en el Punto Décimo, Inciso 10.2 del Acta 28-2011 de fecha 19 de septiembre de 2011, firmo y sello la presente para el trámite correspondiente de graduación de Pregrado.

Atentamente,

"Id y Enseñad a Todos"

Mtro. Lilia Susana Beltrán Paiz
Asesor(a)

Lilia Susana Beltrán Paiz
Ingeniera en Sistemas y
Magister en Informática
Colegiado No. 6754

UNIVERSIDAD DE SAN CARLOS
ESCUELA DE POSTGRADO
FACULTAD DE INGENIERIA
DE GUATEMALA

Mtro. Mario Renato Escobedo Martinez
Coordinador(a) de Maestría

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERIA
DIRECCIÓN
ESCUELA DE POSTGRADO

Mtro. Edgar Dario Alvaréz Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería

UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÍA

EEP-EIME-0147-2022

El Director de la Escuela De Ingenieria Mecanica Electrica de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del Asesor, el visto bueno del Coordinador y Director de la Escuela de Estudios de Postgrado, del Diseño de Investigación en la modalidad Estudios de Pregrado y Postgrado titulado: **DISEÑO DE UNA RED NEURONAL PARA LA ESTIMACIÓN DE LA PROBABILIDAD DE ÉXITO DE UN NEGOCIO EN UNA EMPRESA DE DISTRIBUCIÓN DE PRODUCTOS.**, presentado por el estudiante universitario **Juan Carlos Martini Palma**, procedo con el Aval del mismo, ya que cumple con los requisitos normados por la Facultad de Ingeniería en esta modalidad.

ID Y ENSEÑAD A TODOS

Ing. Armando Alonso Rivera Carrillo
Director
Escuela De Ingenieria Mecanica Electrica

Guatemala, enero de 2022

Decanato
Facultad de Ingeniería
24189101- 24189102
secretariadecanato@ingenieria.usac.edu.gt

LNG.DECANATO.OI.542.2022

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería Mecánica Eléctrica, al Trabajo de Graduación titulado: **DISEÑO DE INVESTIGACIÓN DE UNA HERRAMIENTA PARA LA ESTIMACIÓN DE LA PROBABILIDAD DE ÉXITO DE UN NEGOCIO EN UNA EMPRESA DE DISTRIBUCIÓN DE PRODUCTOS UTILIZANDO MÉTODOS DE APRENDIZAJE AUTOMÁTICO,** presentado por: **Juan Carlos Martini Palma,** después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANA
FACULTAD DE INGENIERÍA

Inga. Aurelia Anabela Cordova Estrada
Decana

Guatemala, julio de 2022

AACE/gaoc

ACTO QUE DEDICO A:

Mis padres	Juan Carlos Martini y Hilda Palma, por su apoyo y su amor incondicional.
Mis hermanas	Por apoyarme a lo largo de mi vida.
Mi familia	Por siempre estar conmigo en los momentos difíciles.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por permitirme formarme como persona y profesional.
Mis amigos	Del departamento de matemática y compañeros de universidad, por todos los buenos momentos y el apoyo que brindaron durante la carrera.
Catedráticos	Por la formación brindada y el libre intercambio de conocimientos.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
$\boldsymbol{sgn}(\boldsymbol{x})$	Función signo de x
$\|\|\boldsymbol{w}\|\|$	Magnitud del vector w
$\boldsymbol{w}^T$	Transpuesta del vector w

GLOSARIO

Algoritmo Es un conjunto de instrucciones o reglas definidas y no ambiguas, ordenadas y finitas que permite, típicamente, solucionar un problema.

Hiperplano Un subespacio con una dimensión n-1 dentro de un espacio de dimensión n.

Kivy Librería de Python utilizada para crear apps y herramientas con una interfaz de usuario.

Scikit-learn Librería de Python para la creación de herramientas de aprendizaje automático.

Sistema binario Sistema en donde solo se utilizan dos dígitos.

Tensorflow Librería de Python para la creación de herramientas de aprendizaje automático e inteligencia artificial.

RESUMEN

El presente trabajo plantea el diseño de una herramienta para la estimación de la probabilidad de éxito de un negocio en una empresa de distribución de productos utilizando métodos de aprendizaje automático.

Para abordar el diseño de la herramienta se recopilaron datos de los negocios anteriores de la empresa, clasificándolos como “exitoso” y “no exitoso”. Se seleccionaron tres algoritmos de aprendizaje automático diferentes para la clasificación de datos, se compararon los resultados de cada uno y se seleccionó el más confiable para la creación de la herramienta.

Con esta propuesta de diseño se pretende reducir el tiempo de selección de un nuevo negocio e incrementar la probabilidad de que este negocio sea rentable para la empresa.

1. INTRODUCCIÓN

Uno de los retos más grandes en la actualidad para las empresas es la digitalización y el uso de las nuevas tecnologías para optimizar procesos que anteriormente necesitaban de mucho tiempo y recurso humano. Esta investigación propone el uso de técnicas de aprendizaje automático para optimizar el proceso de estimación de la probabilidad de éxito de un negocio en una empresa de distribución de productos.

Se espera que esta herramienta sea capaz de brindar resultados confiables y válidos, lo que sería de mucha utilidad para la empresa y los trabajadores de la misma. Los resultados brindados por esta investigación optimizarían el proceso de selección de ofertas de negocios para la empresa.

La herramienta deberá de ser capaz de calcular la probabilidad de éxito de un negocio en una empresa de distribución de productos, teniendo como entrada los datos relevantes del negocio. Este cálculo será de gran ayuda para optimizar el proceso de selección de negocios, lo que podrá brindar crecimiento a la empresa.

Las líneas de investigación de la maestría sobre las cuales se realiza este proyecto son análisis de datos y la minería de datos, por lo que este se considera un proyecto de optimización de procesos.

En la actualidad, se ha explorado la posibilidad de utilizar el aprendizaje automático para calcular la probabilidad de fracaso de empresas que forman parte de la bolsa de valores, lo que ayuda a inversionistas a invertir su dinero de

una forma más eficiente. En esta investigación se espera poder realizar un cálculo similar aplicado a las ofertas de negocio recibidas por una empresa particular.

En el informe final se explicarán las técnicas utilizadas para crear la herramienta, así como el diseño del modelo de aprendizaje de máquinas escogido para calcular la probabilidad. También se incluirá la metodología utilizada para la recolección de datos.

2. ANTECEDENTES

La primera fuente consultada es el trabajo publicado por Lu Wang y Chong Wu (2017) titulado: Predicción de fracasos comerciales basada en un conjunto selectivo de dos etapas con un algoritmo de aprendizaje múltiple y un mapa autoorganizado difuso basado en el *kernel*, el cual es parte de la revista científica *Knowledge-Based Systems* publicado por ELSEVIER.

El proyecto tiene como objetivo crear una red neuronal con el propósito de predecir el fracaso de una empresa. Para esto utiliza la combinación de dos técnicas de aprendizaje automático, así como comparar los resultados obtenidos con resultados obtenidos por otras técnicas de aprendizaje.

El trabajo detalla los variables utilizadas para el estudio de la predicción de fracaso de las empresas y muestra cómo fueron utilizadas para calcular los resltados utilizando las dos tecnicas de aprendizaje automático. También comparan los resultados obtenidos con resultados obtenidos en otros trabajos que utilizan diferentes técnicas, resaltando las más acertadas.

Este ofrece información de las variables utilizadas para el estudio, que son de mucha utilidad para la investigación, ya que los cálculos realizados son muy similares aa los que se pretenden realizar en la investigación. Esto brinda una guía a las variables a escoger para generar la base de datos a utilizar.

La segunda fuente es el trabajo titulado: Aprendizaje automático: una revisión de la clasificación binaria, realizado por Kumari, R., & Srivastava, S. K.

(2017) publicado en el *International Journal of Computer Applications* publicado por Foundation of Computer Science.

El objetivo es enumerar y comparar varios estudios de aprendizaje automático aplicados a problemas de clasificación binaria y ver la evolución de los trabajos a través del tiempo. Se encontró que algunos de los algoritmos que brindan los mejores resultados son los de *random forest* y las máquinas de vectores de soporte.

Se utilizarán durante el proceso, los modelos de *random forest*, máquinas de vectores de soporte y redes neuronales probabilísticas. Estos algoritmos han dado muy buenos resultados en problemas de clasificación binaria a través del tiempo.

La tercera fuente es el estudio titulado: Algoritmos de aprendizaje automático supervisado: clasificación y comparación, realizado por Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., y Akinjobi, J. (2017) publicado en *International Journal of Computer Trends & Technology* publicado por Seventh Sense Research Group.

El trabajo tenía como objetivo comparar y clasificar varios algoritmos de aprendizaje automático utilizando un conjunto de datos complejo de varios datos y un conjunto de datos simple con pocos datos.

Los resultados de la investigación indican que los algoritmos con más aciertos con el conjunto de datos complejo son: *random forest*, SVM, redes neuronales y el clasificador de bayes ingenuo. De esta manera se trabajará con tres de los cuatro algoritmos mencionados.

La cuarta fuente es el trabajo de Tadaaki Hosaka (2018) titulado: Predicción de bancarrota utilizando relaciones financieras en imágenes y redes neuronales convolucionales, el cual fue publicado en *Expert Systems With Applications* publicado por ELSEVIER.

El objetivo de este trabajo es crear una red neuronal capaz de predecir si una empresa declara bancarrota transformando los datos en imágenes para entrenar la red neuronal. También, se comparó los resultados obtenidos por la red neuronal con los resultados obtenidos por otros métodos.

En esta investigación, Hosaka fue capaz de crear una red neuronal capaz de predecir si una empresa declararía bancarrota, demostró que es posible predecir esto con una confiabilidad mayor a otros métodos. También, se llegó a la conclusión, que redes entrenadas con datos creados por el investigador no se comportan de la misma manera que una red entrenada con datos reales.

El trabajo indica que es posible crear una red neuronal capaz de predecir si una empresa declararía bancarrota con una alta confiabilidad, lo que apoya la hipótesis de esta investigación. Por otra parte, se ve que, si se quiere tener mejores resultados, la red debe ser entrenada con datos reales y no creados artificialmente.

Una quinta investigación es la realizada por Ranadeva Jayasekera (2018), titulada: Predicción del fracaso de una empresa: el pasado, presente y direcciones prometedoras para el futuro, publicada en *International Review of Financial Analysis*, publicado por ELSEVIER.

El objetivo del trabajo de Jayasekera fue el de recopilar los trabajos y textos más relevantes en la predicción de fracaso de empresas y a su vez hablar

del futuro de esta área de estudios. También presenta un modelo con mucho potencial para el futuro de estos cálculos.

El informe detalla la historia de los estudios relacionados empezando con un modelo presentado en un estudio de 1930, llegando hasta el presente con un modelo presentado por Jayasekera, que pretende mejorar unos aspectos de los modelos creados a través de los años.

Este estudio es de gran utilidad ya que se obtiene una gran cantidad de información acerca de los modelos creados a través del tiempo. Esto ayuda a la generación del modelo que se pretende crear en esta investigación, así mismo a ver las variables más pertinentes para estos cálculos.

La sexta fuente es el estudio titulado: Predicción del rendimiento en los procesos de negocio utilizando redes neuronales profundas, realizado por Gyunam Park y Minseok Song (2019), el cual fue publicado en *Decision Support Systems* este siendo publicado por ELSEVIER.

El trabajo tenía como objetivo crear un método para predecir el rendimiento futuro de procesos de distintas empresas utilizando una red neuronal de aprendizaje profundo, entrenada con datos históricos de los procesos. Otro de sus objetivos fue el de aplicar la red neuronal en empresas existentes para estudiar su confiabilidad.

El estudio muestra que es posible predecir el rendimiento futuro de procedimientos realizados por una compañía por medio de una red neuronal de aprendizaje profundo entrenada con datos de los procesos anteriores. Los resultados obtenidos muestran que este método genera mejores resultados que los métodos utilizados generalmente por las empresas.

Los resultados de esta investigación reafirman la factibilidad del presente estudio, lo que da una base para esperar que el estudio genere buenos resultados. También confirma que los resultados pueden ser mejores que los obtenidos por los métodos ya establecidos, lo que puede ser beneficioso para la empresa.

La séptima fuente es el trabajo realizado por Ebenuwa, S. H., Sharif, M. S., Alazab, M., y Al-Nemrat, A. (2019) titulado: Técnicas de selección de atributos de clasificación de varianza para problemas de clasificación binaria en datos desequilibrados, publicado en IEEE Access.

El objetivo de la investigación fue presentar una nueva técnica de selección de atributos para un conjunto de datos desproporcionados en problemas de clasificación binaria. Se encontró un incremento apreciable en la precisión de varios algoritmos de clasificación utilizando esta nueva técnica de selección comparada con otras técnicas utilizadas usualmente.

Esta nueva técnica de selección de atributos del conjunto de datos se tomará en cuenta al momento de la selección de datos para aumentar la precisión de los distintos métodos de aprendizaje automático utilizados en esta investigación.

La octava fuente es la investigación realizada por Jiejie Zhao, Bowen Du, Leilei Sun, Weifeng Lv, Yanchi Liu y Hui Xiong (2020), titulada: Aprendizaje multitarea profundo con atención relacional para la predicción del éxito de una empresa, publicada en la revista *Pattern Recognition*, publicada por ELSEVIER.

El trabajo tuvo como objetivo la creación de un esquema de aprendizaje automático multitarea, que fuera capaz de encontrar relaciones entre procesos y

así poder generar un nuevo entendimiento de los mismos y su relación con el éxito de la empresa. También, se dieron a la tarea de utilizar el esquema en empresas reales, para estudiar el desempeño del mismo.

La investigación detalla la creación de un esquema de aprendizaje automático multitarea utilizado específicamente en el calcula de la probabilidad de éxito de una cadena de tiendas, basadas en la ubicación de la misma. Los investigadores utilizaron una gran cantidad de variables para encontrar la mejor ubicación para la tienda. Estos basados en las salidas observadas del esquema de aprendizaje automático profundo de múltiples tareas. Los resultados obtenidos fueron muy buenos, mejores que los obtenidos por los procesos utilizados por las empresas.

Este estudio indica que los resultados obtenidos por métodos de aprendizaje automáticos pueden ser tomados en cuenta al momento de tomar decisiones empresariales. Estos resultados pueden traer grandes beneficios a una gran cantidad de empresas, que es lo que se busca realizar en la presente investigación.

La novena fuente es el trabajo realizado por Santoso, L., Singh, B., Rajest, S., Regin, R., y Kadhim, K. (2020) titulado: Un enfoque de programación genética para el problema de clasificación binaria, publicado en EAI Endorsed Transactions on Energy Web.

El trabajo tenía como objetivo utilizar un algoritmo genético para resolver varios problemas de clasificación binaria utilizando diferentes tipos de conjuntos de datos. Los resultados de la investigación fueron buenos, pero se necesitaron muchas generaciones del algoritmo para tener un resultado aceptable lo que tomaba mucho tiempo y recursos computacionales.

Estos resultados indican que los algoritmos de aprendizaje automático son buenas herramientas para resolver problemas de clasificación binaria, pero es necesario tomar en consideración los recursos necesarios para el entrenamiento de los algoritmos.

La décima fuente se titula: Las ventajas del coeficiente de correlación de Matthews (MCC) sobre la puntuación F1 y la precisión en la evaluación de la clasificación binaria, realizada por Chicco, D., y Jurman, G. (2020) publicado en BMC genomics.

La investigación tenía como objetivo comprar distintos indicadores del rendimiento en problemas de clasificación binaria. El estudio comparó el coeficiente de correlación de Matthews con el puntaje F1 y la precisión.

Los resultados obtenidos por la investigación indican que el coeficiente de correlación de Matthews es el mejor indicador, ya que toma en cuenta los cuatro posibles resultados de las pruebas y la proporción de casos positivos y negativos del conjunto de datos. Por lo que el indicador principal que se utilizará en esta investigación será el coeficiente de correlación de Matthews.

3. PLANTEAMIENTO DEL PROBLEMA

3.1. Contexto general

Esta empresa recibe varias ofertas de posibles negocios de varios proveedores por lo que utiliza bastantes recursos en el estudio y selección de estas ofertas. Varios empleados de diferentes áreas realizan diferentes estudios para encontrar la viabilidad del negocio y si será beneficioso para la empresa. El proceso de selección toma un tiempo considerable para llegar a una conclusión y esta empresa recibe varias ofertas cada año.

3.2. Descripción del problema

Al momento de tener demasiadas ofertas de negocio la carga laboral de los empleados se ve incrementada, sumándole a sus demás responsabilidades, esto puede resultar en ofertas evaluadas de manera errónea o en ofertas rechazadas sin llevar a cabo el debido proceso. Esto podría resultar en el rechazo de ofertas que podrían ser de gran beneficio para la empresa, lo que podría frenar el crecimiento de la empresa y llevar a la pérdida de negocios rentables.

Este proceso tarda un tiempo considerable lo que puede atrasar otros procesos importantes de la empresa, lo que disminuye la eficiencia de la empresa y puede llevar a pérdidas. El tiempo que lleva realizar esto proceso se podría utilizar para otras tareas importantes que aumenten la eficiencia y la calidad del trabajo de los empleados.

3.3. Formulación del problema

Proceso ineficiente en la selección de nuevos negocios.

3.3.1. Pregunta central

Esto lleva a plantear la pregunta principal de este estudio: ¿Como calcular la probabilidad de éxito de un negocio en una empresa de distribución de productos utilizando distintos métodos de aprendizaje automático entrenados con datos de negocios anteriores?

3.3.2. Preguntas auxiliares

Para responder a esta interrogante se deberán contestar las siguientes preguntas auxiliares:

- ¿Cuáles son los atributos del conjunto de datos con más peso al momento de calcular la probabilidad de éxito en un negocio?
- ¿Cuál es el grado de confianza de los resultados generados por los distintos métodos de aprendizaje automático?
- ¿Cuál de los dos métodos, la herramienta y el método utilizado por la empresa, es más preciso y eficiente para calcular la probabilidad de éxito de un negocio?

3.4. Delimitación del problema

El problema se resolverá creando una herramienta entrenada con datos de los negocios anteriores de la empresa. Se dividirá el conjunto de datos en dos

sets: un set de aprendizaje y otro de prueba para poder evaluar los resultados del aprendizaje.

4. JUSTIFICACIÓN

El desarrollo de este proyecto se justifica en las líneas de investigación del análisis de datos y minería de datos de la Maestría en Ingeniería para la Industria con Especialización en Ciencias de la Computación, ya que esta implica el uso de técnicas de aprendizaje automáticas y la creación de una base de datos de los negocios anteriores de la empresa.

Por medio de este, se mostrará que es posible crear una herramienta utilizando técnicas de aprendizaje automático que sea capaz de calcular la probabilidad de éxito de un negocio. Esto beneficiará a la empresa en cuestión ya que se podrá agilizar el proceso de selección de negocios rentables y podrá ser una herramienta indispensable para el crecimiento de dicha empresa.

Al concluir esta investigación, se contará con un modelo de aprendizaje de máquinas que será capaz de calcular la probabilidad de éxito de un negocio en una empresa de distribución teniendo como entrada los datos relevantes del negocio. También se creará un proceso más eficiente y simple para obtener los datos importantes de las ofertas de negocio por medio de herramientas de obtención y análisis de datos.

El personal de la empresa será capaz de calcular la probabilidad de éxito de un negocio al momento de recibir ofertas para nuevos negocios, lo que optimizará estos procesos en la empresa. En el pasado estos procesos requerían del trabajo de varias personas de diferentes áreas para poder llegar a una conclusión de la probabilidad de éxito del negocio. Esta investigación disminuirá

el recurso humano requerido para estos cálculos, lo que será beneficioso para la empresa.

La investigación tiene la posibilidad de disminuir la carga laboral de los empleados de la empresa, lo que podría llevar a una reducción de los niveles de estrés del personal. Por otro lado, la toma de decisiones basadas en los resultados de la herramienta será más objetiva, ya que solo se toman en cuenta datos verdaderos y comprobables, lo que hará de este un proceso más justo y objetivo.

5. OBJETIVOS

5.1. General

Crear una herramienta capaz de calcular la probabilidad de éxito de un negocio en una empresa de distribución de productos.

5.2. Específicos

- Identificar los atributos más importantes del conjunto de datos en el cálculo de la probabilidad de éxito de un negocio en una empresa de distribución de productos.

- Calcular la confiabilidad de los resultados obtenidos a través de los distintos modelos de aprendizaje automático.

- Comparar los recursos utilizados y los resultados obtenidos con la herramienta con los resultados obtenidos por la empresa.

6. NECESIDADES POR CUBRIR Y ESQUEMA DE SOLUCIÓN

Una de las decisiones más importantes para una empresa es seleccionar la oferta de negocio que más le beneficie y que tenga una mayor probabilidad de ser exitosa, por lo que siempre se encuentran en la búsqueda de optimizar este proceso. En la actualidad, esta empresa no cuenta con un proceso automático para realizar estos estudios por lo que se utiliza una gran cantidad de tiempo y recurso humano. Esta investigación busca presentar una alternativa más simple y eficiente para realizar estos procesos con la ayuda del análisis de datos y el aprendizaje de máquinas. Esta empresa no cuenta con un proceso automático para realizar estos estudios.

Muchas empresas están acostumbradas a hacer procesos de la misma manera por varios años, lo que hace complicado el presentar una forma nueva de realizar procesos ya establecidos. También es bastante común que estas no confíen en los resultados obtenidos por medio de procesos automáticos, la mayoría de las veces confían más en resultados obtenidos por personas. Por esta razón, es de alta importancia crear una herramienta que brinde resultados con una alta confiabilidad para poder demostrar que el producto de esta investigación es una metodología válida.

Estos procesos no optimizados reducen la eficiencia de la compañía y aumentan la carga laboral de los empleados de varias áreas, lo que puede disminuir las ganancias de la compañía y puede afectar el rendimiento de los trabajadores. El resultado de esta investigación pretende optimizar estos procesos y disminuir la carga laboral de los empleados lo que lo hace valioso para la empresa.

Se utilizarán varios modelos de aprendizaje automático, estos siendo entrenada con un conjunto de datos de los negocios seleccionados en años anteriores por la empresa utilizando el proceso convencional.

7. MARCO TEÓRICO

7.1. Problemas de clasificación binaria

Los problemas de clasificación binaria se pueden definir como el proceso de clasificar algún elemento en base a dos clases predefinidas (Kumari, Roshan, Srivastava y Saurabh, 2017). Uno de los ejemplos más conocidos es el de la clasificación de personas en una de dos clases: “enfermas” y “no enfermas”. Las pruebas de clasificación binaria pueden dar cuatro resultados:

- Verdadero positivo (TP): personas enfermas clasificadas como “enfermas”
- Falso positivo (FP): personas sanas clasificadas como “enfermas”
- Verdadero negativo (TN): personas sanas clasificadas como “no enfermas”
- Falso negativo (FN): personas enfermas clasificadas como “no enfermas”

7.2. Modelos de aprendizaje automático para problemas de clasificación binaria

A lo largo de la historia los problemas de clasificación binaria aplicados a la predicción de éxito de un negocio han sido solucionados de dos formas distintas. El primer método es por medio de métodos estadísticos como lo es el análisis de discriminantes múltiples Altman (1968) o por medio de modelos estocásticos como logit (Ohlson, 1980). El método más reciente es la utilización de modelos de aprendizaje automático en el cual se han utilizado varios métodos, como el árbol de decisión (Frydman, Altman y Kao, 1985) y las máquinas de vectores de soporte (SVM). (Min y Lee, 2005)

En este trabajo de investigación se utilizarán los siguientes modelos de aprendizaje automático: máquinas de vectores de soporte, *random forest* y redes neuronales convolucionales.

7.3. Máquinas de vectores de soporte (SVM)

Las máquinas de vectores de soporte son algoritmos de aprendizaje automático supervisado utilizado principalmente en problemas de clasificación y regresión. Fueron mencionadas por primera vez en 1995 (Vapnik & Cortes, 1995) donde se presenta la idea de un modelo que mapea los vectores de entrada a un espacio de mayor dimensión utilizando un mapeo no linear escogido previamente.

Con un conjunto de datos con la forma siguiente:

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2) \dots (\boldsymbol{x}_n, y_n)$$

Donde $\boldsymbol{x}_i$ es un vector p-dimensional e y_i puede tener un valor que puede ser 1 o -1 el cual indica a clase pertenece el vector. Se requiere encontrar el hiperplano que divide el grupo de vectores donde $y_i = 1$ del grupo de vectores donde $y_i = -1$ con la mayor separación posible.

Un hiperplano puede definirse en este contexto como el conjunto de puntos que cumplen la siguiente igualdad:

$$\boldsymbol{w}^T \boldsymbol{x} - b = 0$$

Donde $\boldsymbol{w}$ es el vector normal al hiperplano y b es la separación del origen.

Si se tiene un conjunto de datos estandarizado con $y_i = 1, -1$ se pueden definir dos hiperplanos que separan los dos diferentes grupos de datos utilizando las siguientes ecuaciones:

$\boldsymbol{w}^T\boldsymbol{x} - b = 1\ (todo\ lo\ que\ este\ arriba\ de\ este\ limite\ es\ del\ grupo\ donde\ y_i = 1)$
$\boldsymbol{w}^T\boldsymbol{x} - b = -1 (todo\ lo\ que\ este\ arriba\ de\ este\ limite\ es\ del\ grupo\ donde\ y_i = -1)$

Calculando la distancia entre los dos hiperplanos se obtiene que la distancia es:

$$d = \frac{2}{\|\boldsymbol{w}\|}$$

Por lo que para poder maximizar la separación entre los hiperplanos se debe minimizar el término $\|\boldsymbol{w}\|$.

Aplicando la siguiente restricción para que los datos no puedan estar dentro del hiperplano:

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i - b) \geq 1\ para\ 1 \leq i \leq n$$

Se obtiene el problema de optimización que se pretende resolver:

$$Minimizar\ \|\boldsymbol{w}\|\ bajo\ y_i(\boldsymbol{w}^T\boldsymbol{x}_i - b) \geq 1\ para\ 1 \leq i \leq n$$

El $\boldsymbol{w}$ y b que solucionen este problema de optimización determinara el clasificador que se utilizara para futuros datos.

Las máquinas de vectores de soporte han tenido buenos resultados en problemas de predicción de éxito de un negocio (Keplac y Hampel, 2016) por lo que se creara un modelo de SVM en este trabajo de investigación para calcular la probabilidad de éxito de un negocio en una empresa de distribución de productos.

Algunos de los posibles inconvenientes de las máquinas de vectores de soporte son los siguientes:

- Requiere que todo el conjunto de datos este etiquetado para poder aprender correctamente
- Los parámetros que se obtienen de un modelo de SVM son difíciles de interpretar.

7.4. *Random Forest*

Random forest es un algoritmo de aprendizaje automático utilizado ampliamente en problemas de clasificación y regresión. Este algoritmo crea varios de árboles de decisión durante el aprendizaje y da como resultado la clase que fue seleccionada por la mayoría de árboles de decisión.

Se puede decir que este método de entrenamiento fue propuesto por primera vez por Tim Kan Ho (Ho, 1995) cuando se utiliza el método de subespacios aleatorios en los que se seleccionaban características aleatorias de la población para entrenar a varios árboles de decisión.

Dado un conjunto de datos $X = x_1, x_2, \dots x_n$ con clases $Y = y_1, y_2, \dots y_n$ se crean B muestras aleatorias con remplazo $X_1, X_2, \dots X_B$ y se entrena un árbol de decisión f_i con cada muestra. Después del entrenamiento se pueden clasificar

datos nunca antes vistos por el modelo en base a la clase que fue escogida por la mayor cantidad de árboles de decisión.

Con $y_i = 1, -1$ y un dato nunca antes visto por el modelo $\acute{x}$ se puede clasificar por medio de:

$$\acute{y} = sgn\left(\sum_{1}^{B} f_i(\acute{x})\right)$$

El método de *random forest* tiende a dar muy buenos resultados con conjuntos de datos no balanceados y con ciertos parámetros faltantes. (Lin, Wu, Lin, Wen y Li, 2017)

Posibles desventajas de utilizar el algoritmo de *random forest* son las siguientes:

- Es difícil interpretar los resultados por la gran cantidad de árboles que se crean para la clasificación.
- La cantidad de árboles de decisión creados pueden aumentar el tiempo de aprendizaje comparado con otros algoritmos similares.

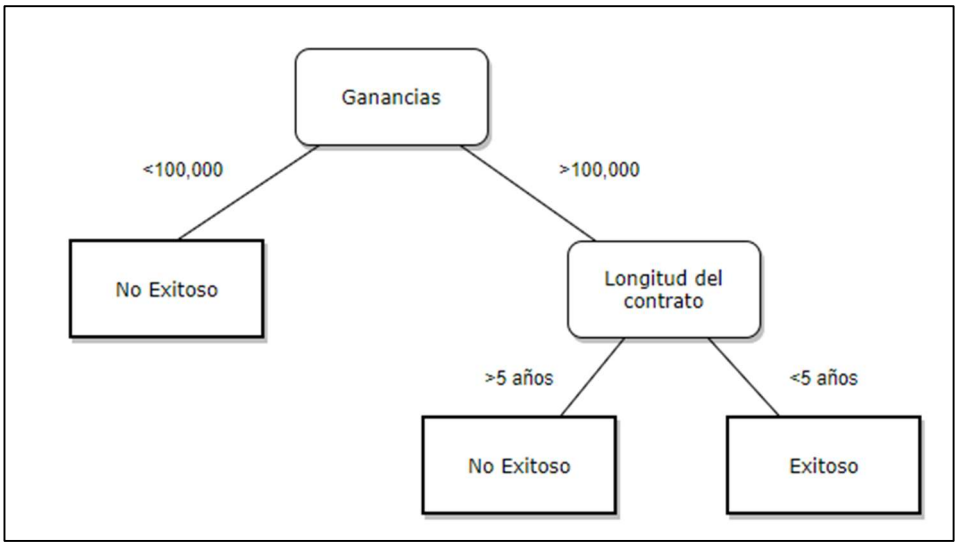
7.4.1. Árboles de decisión

Los árboles de decisión son modelos de predicción utilizados en estadística y el aprendizaje automático. Existen dos clases de árboles de decisión: los árboles regresivos, en donde resultado obtenido puede ser continuo y árboles clasificatorios, en donde el valor de la salida solo puede tomar valores discretos. Se usarán árboles clasificatorios en este trabajo de investigación.

En los árboles clasificatorios las hojas del árbol representan las etiquetas de las distintas clases y las ramas representan los parámetros que conducen a dichas etiquetas.

Un ejemplo de un árbol de decisión clasificatorio es el siguiente:

Figura 1. **Ejemplo de árbol de decisión**



Fuente: elaboración propia, empleando draw.io.

7.5. Redes neuronales artificiales (ANN)

Las ANN son sistemas computacionales inspirados por las redes neuronales biológicas que se encuentran en los cerebros de varios animales. Estas redes están formadas por varios elementos:

- Neuronas artificiales: cada neurona puede tener varias entradas y una sola salida que puede ser entregada a varias otras neuronas. Para encontrar la

salida de una neurona se de tomar la suma ponderada de todas las entradas y luego se le aplica una función de activación al resultado de la suma.

- Conectores y pesos: los conectores proporcionan la salida de una neurona calculada anteriormente a la entrada de otra multiplicada por un peso.

- Organización de la red: usualmente las neuronas son organizadas en varias capas, las neuronas que pertenecen una capa se conectan solamente con neuronas de capas aledañas. Se tiene la capa de entrada, donde se brinda la entrada de los datos a estudiar, y la capa de salida, donde se obtiene la salida de la red neuronal, se pueden tener cero o más capas entre la capa de entrada y salida.

- Hiperparámetro: es un parámetro constante cuyo valor se escoge antes de proceso de aprendizaje. Ejemplos de hiperparámetros pueden ser la cantidad de capas o el ritmo de aprendizaje.

Las redes neuronales artificiales pueden tener varias aplicaciones en los negocios ya que pueden ayudar a predecir o a la toma de decisiones (Li, 1994). Algunas de las desventajas en la aplicación de las redes neuronales artificiales son las siguientes:

- Las redes neuronales solo son aplicables a ciertos tipos de problemas, por lo que se debe tener criterio al momento de presentarlos como una solución.

- Se debe tener una gran cantidad de datos para poder diseñar una buena red neuronal.

- Es difícil interpretar los resultados de una red neuronal ya que su funcionamiento es muy complejo.

Existe una gran variedad de tipos de redes neuronales con diferentes ventajas y desventajas, en este trabajo de investigación se utilizará una red neuronal probabilística.

7.5.1. Redes neuronales probabilísticas

Este tipo red neuronal creado en 1995 (Specht, 1995) es ampliamente utilizado en problemas de reconocimiento de patrones y de clasificación. Las redes neuronales probabilísticas son redes neuronales unidireccionales las cuales están compuestas por cuatro capas:

- Capa de entrada: cada neurona en esta capa corresponde a una variable predictiva del caso de prueba el cual se desea clasificar.
- Capa de patrón: cada neurona de esta capa corresponde a un caso del set de entrenamiento, estas neuronas calculan la distancia entre el caso de prueba y el caso en cada neurona.
- Capa de suma: cada neurona de esta capa corresponde a las clases en las cuales se desea clasificar los datos, en esta capa se realiza una suma ponderada de los resultados de la capa anterior.
- Capa de salida: esta capa compara los resultados obtenidos por la capa de suma y entrega como resultado la clase con la mayor suma.

Algunas ventajas de estas redes sobre otros tipos de redes neuronales son las siguientes:

- Las redes neuronales probabilísticas tienden a ser más rápidas y eficientes en problemas de clasificación que otros tipos de redes neuronales.
- Las redes neuronales probabilísticas son relativamente insensibles a datos atípicos.
- Las redes neuronales probabilísticas se asemejan a la clasificación optima de Bayes.

7.6. Coeficiente de correlación de Matthews

El coeficiente de correlación de Matthews es una medida utilizada en el aprendizaje automático para medir la calidad de una clasificación binaria. Fue creado por el bioquímico Brian Matthews en 1975 (Matthews, 1975) y desde entonces ha sido utilizado en varios problemas de clasificación binaria. Es posible calcular el coeficiente utilizado la matriz de confusión con la formula:

Tabla I. **Matriz de confusión**

Clase predicha / **Clase actual**	**Positivo**	**Negativo**
Positivo	TP (verdadero positivo)	FN (falso negativo)
Negativo	FP (falso positivo)	TN (verdadero negativo)

Fuente: elaboración propia, empleando Excel.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

El coeficiente puede devolver valores entre -1 y 1, siendo 1 una predicción perfecta, 0 una predicción no mejor a una clasificación aleatoria y -1 una predicción completamente errónea.

Se recomienda utilizar el coeficiente de correlación de Matthews sobre otras medidas como la precisión y el valor F1 ya que el coeficiente de Matthews si toma en cuenta las relaciones de los cuatro resultados posibles de la clasificación: verdadero positivo, falso positivo, verdadero negativo y falso negativo por lo que el modelo de clasificación debe tener un alto grado de aciertos en los casos positivos y negativos para poder tener una buena puntuación.

8. PROPUESTA DE ÍNDICE DE CONTENIDOS

9. METODOLOGÍA

9.1. Características del estudio

El enfoque del estudio es cuantitativo, ya que se obtendrán datos estadísticos de varios modelos de aprendizaje automático con la tarea de obtener la probabilidad de éxito de un negocio en una empresa de distribución de productos y con los datos obtenidos encontrar el modelo más adecuado para la tarea. También se diseñará una herramienta capaz de calcular la probabilidad de éxito de un negocio con el modelo más preciso, que será entregada a la empresa en cuestión.

El alcance es descriptivo, ya que con los datos generados se obtendrá el modelo de aprendizaje más apropiado y se encontrarán los elementos que tienen mayor influencia al momento de generar la probabilidad de éxito de un negocio, lo que ayudará a comprender de mejor manera el proceso.

El diseño de la investigación será no experimental, ya que no se manipularán variables independientes.

9.2. Unidades de análisis

La población que se estudiará será una recopilación de los negocios aceptados por la empresa en los últimos 10 años, que tendrán etiquetas de “exitoso” y “no exitoso”. La población se dividirá en dos muestras, una muestra será utilizada para el aprendizaje del modelo y la otra muestra será utilizada para

validar el aprendizaje de los modelos. Las muestras serán obtenidas por medio de un muestreo aleatorio simple.

9.3. Variables

Las variables del estudio se presentan a continuación:

Tabla II. **Variables del estudio**

Variable	Definición conceptual	Definición operacional
Modelo de aprendizaje automático	El tipo de modelo de aprendizaje automático con sus distintos atributos.	Se entrenarán varios modelos de aprendizaje automático con el mismo conjunto de datos.
Tiempo de aprendizaje promedio	El tiempo que toma en entrenar un modelo especifico.	El tiempo de aprendizaje promedio se medirá en minutos.
Tiempo de inferencia promedio	El tiempo que toma un modelo especifico en devolver una predicción dado un conjunto de datos nunca antes visto por el modelo.	El tiempo de inferencia promedio se medirá en milisegundos.

Fuente: elaboración propia empleando Excel.

9.4. Fases del estudio

El estudio consta de seis fases.

9.4.1. Revisión de la literatura

Se obtendrá el conocimiento teórico necesario para poder realizar los experimentos requeridos por la investigación. Se revisarán principalmente trabajos científicos relacionados al aprendizaje automático y sus aplicaciones en los procesos corporativos.

9.4.2. Recolección y clasificación de los datos

Esta fase constará en la recolección de los datos de los negocios anteriores de la empresa que serán brindados por la misma. Se realizará una depuración de los datos obtenidos con el objetivo de remover los parámetros que no sean relevantes para el estudio. Adicionalmente se agregarán etiquetas de “exitoso” y “no exitoso” a cada negocio en el conjunto de datos y se crearán dos muestras por medio del muestreo aleatorio simple.

9.4.3. Elaboración de los modelos de aprendizaje automático

Se elaborarán los modelos de aprendizaje automático seleccionados en base a lo estudiado en la literatura. Se crearán de manera que el cambio de hiperparámetros se pueda realizar de una forma rápida y eficiente.

9.4.4. Entrenamiento de los diferentes modelos de aprendizaje

En esta fase, se entrenarán los modelos utilizando la muestra previamente obtenida para el entrenamiento. Los modelos serán evaluados utilizando el coeficiente de correlación de Matthews y los hiperparámetros de estos serán modificados para obtener el mejor coeficiente posible.

9.4.5. Diseño de la herramienta para calcular la probabilidad de éxito de un negocio

Se desarrollará una herramienta capaz de calcular la probabilidad de éxito de un negocio con el modelo que brinde los mejores resultados. Esta herramienta deberá ser de fácil utilización para que sea utilizada por la empresa en cuestión.

9.4.6. Redacción del informe final

Se realizará un informe describiendo los resultados obtenidos por la investigación.

10. TÉCNICAS DE ANÁLISIS DE INFORMACIÓN

Para llevar a cabo el análisis de la información se utilizarán varias métricas:

La sensibilidad indica la capacidad del estimador para dar como casos exitosos los casos realmente exitosos.

$$sensibilidad = \frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ negativos}$$

La especificidad indica la capacidad del estimador para dar como casos no exitosos los casos realmente no exitosos.

$$especificidad = \frac{verdaderos\ negativos}{verdareos\ negativos + falsos\ positivos}$$

El coeficiente de correlación de Matthews ayuda a calcular la calidad de un proceso de clasificación binaria. Esta métrica será la principal al momento de comparar modelos de aprendizaje.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Donde:

$TP = verdaderos\ positivos$

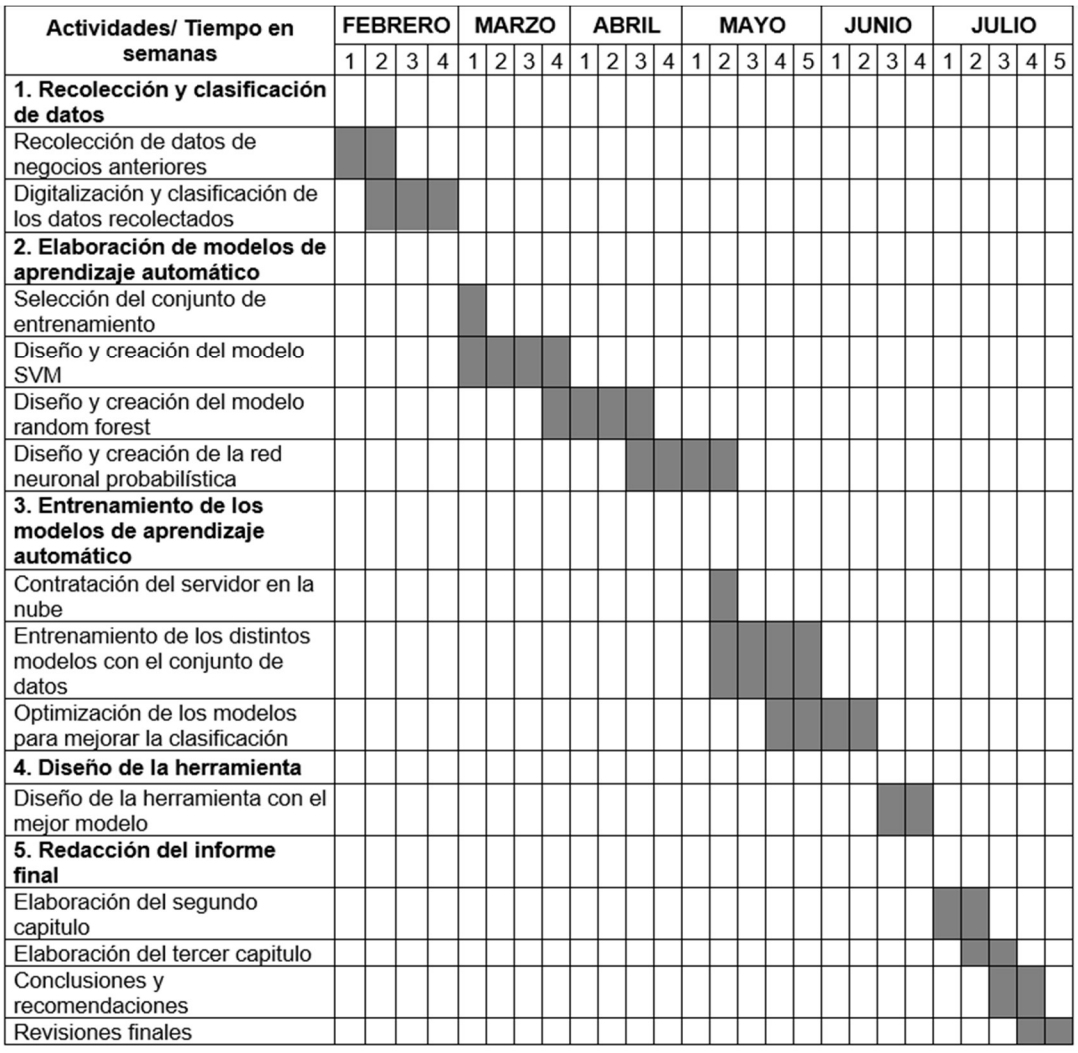
$TN = verdaderos\ negativos$

$FP = falsos\ positivos$

$FN = falsos\ negativos$

11. CRONOGRAMA

Figura 2. **Cronograma**



Fuente: elaboración propia, empleando Microsoft Project.

12. FACTIBILIDAD DEL ESTUDIO

12.1. Factibilidad económica

A continuación, en la tabla III se describen los gastos del estudio.

Tabla III. **Gastos del estudio**

Rubro	**Costo**	**Observación**
Servidor en la nube	Q. 350.00 /mes por 3 meses	Servidor donde se ejecutará el aprendizaje de los modelos
Papelería	Q. 100.00	Cartas, tramites, etc.
Total	Q. 1150.00	

Fuente: elaboración propia empleando Excel.

12.2. Factibilidad operativa

- Recolección de los datos

Los datos serán bridados por una empresa privada de distribución de productos, con información de negocios exitosos y no exitosos de los últimos 10 años. Los datos deberán ser limpiados y etiquetados para su uso en la investigación.

- Entrenamiento de los modelos

Se alquilará un servidor en la nube donde se ejecutarán los distintos modelos de aprendizaje para agilizar el entrenamiento y evitar problemas de disponibilidad de equipo de cómputo.

12.3. Factibilidad técnica

Se utilizará el lenguaje de programación Python para crear los modelos utilizando las librerías:

- *Tensorflow* y *Scikit-learn* para crear los distintos modelos de aprendizaje automático
- *Kivy* para la creación de la herramienta prototipo

Todas las librerías a utilizar son de código abierto.

13. REFERENCIAS

1. Altman, E. (septiembre, 1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance, 23*(4), 589–609.

2. Chicco, D. y Jurman, G. (enero, 2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics, 21*(1), 1-13.

3. Cortes, C. y Vapnik, V. (marzo, 1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

4. Ebenuwa, S. H., Sharif, M. S., Alazab, M. y Al-Nemrat, A. (marzo, 2019). Variance ranking attributes selection techniques for binary classification problem in imbalance data. *IEEE Access, 7*, 24649-24666.

5. Ho, T. (agosto, 1995). Random decision forests. *In Proceedings of 3rd international conference on document analysis and recognition. Vol. 1*. 278-282.

6. Hosaka, T. (julio, 2018). Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Elsevier,* 1-38.

7. Jayasekera, R. (agosto, 2018). Prediction of company failure: Past, present and promising directions for the future. *International Review of Financial Analysis, 55*, 196–208. doi: 10.1016/j.irfa.2017.08.009

8. Klepáč, V. y Hampel, D. (mayo, 2016). Prediction of bankruptcy with SVM classifiers among retail business companies in EU. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 64*(2), 627-634.

9. Kumari, R. y Srivastava, S. (febrero, 2017). Machine Learning: A Review on Binary Classification. *International Journal of Computer Applications, 160*(7), 11-15. Doi. 10.5120/ijca2017913083.

10. Li, E. (noviembre, 1994). Artificial neural networks and their business applications. *Information & Management, 27*(5), 303-313.

11. Lin, W., Wu, Z., Lin, L., Wen, A. y Li, J. (julio, 2017). An ensemble *random forest* algorithm for insurance big data analysis. *Ieee access, 5*, 16568-16575.

12. Marais, M., Patel, J. y Wolfson, M. (1984). The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications. *Journal of Accounting Research, 22*(Suppl.), 87–114.

13. Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure, 405*(2), 442-451.

14. Min, J. y Lee, Y. (mayo, 2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications, 28*(4), 603–614.

15. Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109–131.

16. Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O. y Akinjobi, J. (julio, 2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT), 48*(3), 128-138.

17. Park, G., & Song, M. (2019). Predicting performances in business processes using deep neural networks. Decision Support Systems, 113191. doi: 10.1016/j.dss.2019.113191

18. Santoso, L., Singh, B., Rajest, S., Regin, R., & Kadhim, K. (2020). A Genetic Programming Approach to Binary Classification Problem. EAI Endorsed Transactions on Energy Web, 8(31), e11.

19. Specht, D. F. (1990). Probabilistic neural networks. Neural networks, 3(1), 109-118.

20. Wang, L., & Wu, C. (2017). Business failure prediction based on two-stage selective ensemble with manifold learning algorithm and kernel-based fuzzy self-organizing map. Knowledge-Based Systems, 121, 99–110. doi: 10.1016/j.knosys.2017.01.016

21. Zhao, J., Du, B., Sun, L., Lv, W., Liu, Y., & Xiong, H. (2020). Deep Multi-task Learning with Relational Attention for Business Success Prediction. Pattern Recognition, 107469. doi: 10.1016/j.patcog.2020.107469