Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

DISEÑO DE INVESTIGACIÓN DE UNA HERRAMIENTA PARA EL ANÁLISIS DE DOCUMENTOS DE RECLAMOS EN UNA ASEGURADORA UTILIZANDO MACHINE LEARNING

Marvin Wilfredo Ajcuc Cuzco
Asesorado por el M.A. Ing. Dennis Fernando Palma Ramírez

Guatemala abril de 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

DISEÑO DE INVESTIGACIÓN DE UNA HERRAMIENTA PARA EL ANÁLISIS DE DOCUMENTOS DE RECLAMOS EN UNA ASEGURADORA UTILIZANDO MACHINE LEARNING

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

MARVIN WILFREDO AJCUC CUZCO
ASESORADO POR EL ING. DENNIS FERNANDO PALMA RAMÍREZ

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, ABRIL 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 Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. Pedro Antonio Aguilar Polanco
EXAMINADOR	Ing. Marlon Francisco Orellana López
EXAMINADOR	Ing. Carlos Alfredo Azurdia Morales
EXAMINADOR	Ing. César Augusto Fernández Cáceres
SECRETARIA	Ing. Lesbia Magalí Herrera López

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 EL ANÁLISIS DE DOCUMENTOS DE RECLAMOS EN UNA ASEGURADORA UTILIZANDO MACHINE LEARNING

Tema que me fuera asignado por la Dirección de la Escuela de Estudios de Postgrado, con fecha 1 de diciembre de 2021.

Marvin Wilfredo Ajcuc Cuzco

ESCUELA DE ESTUDIOS DE
POSTGRADO
FACULTAD DE INGENIERÍA

EEPFI-PP-0158-2022
Guatemala, 12 de enero de 2022

Director
null
Escuela De Ingenieria En Sistemas
Presente.

Estimado Ing. Alonzo

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 HERRAMIENTA PARA EL ANÁLISIS DE DOCUMENTOS DE RECLAMOS EN UNA ASEGURADORA UTILIZANDO MACHINE LEARNING**, el cual se enmarca en la línea de investigación: **Análisis de datos - Análisis de datos**, presentado por el estudiante **Marvin Wilfredo Ajcuc Cuzco** carné número **200516326**, 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"

Dennis Fernando Palma Ramírez
Ingeniero en Sistemas de Información y Ciencias de la Computación
Colegiado No. 18,870

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

Mtro. Dennis Fernando Palma Ramírez
Asesor(a)

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

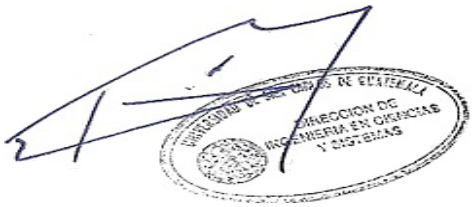
UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÍA

EEP-EICS-0158-2022

El Director de la Escuela De Ingenieria En Sistemas 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 HERRAMIENTA PARA EL ANÁLISIS DE DOCUMENTOS DE RECLAMOS EN UNA ASEGURADORA UTILIZANDO MACHINE LEARNING**, presentado por el estudiante universitario **Marvin Wilfredo Ajcuc Cuzco**, 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. Carlos Gustavo Alonzo
Director
Escuela De Ingenieria En Sistemas

Guatemala, enero de 2022

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

LNG.DECANATO.OI.231.2022

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 en Ciencias y Sistemas, al Trabajo de Graduación titulado: **DISEÑO DE INVESTIGACIÓN DE UNA HERRAMIENTA PARA EL ANÁLISIS DE DOCUMENTOS DE RECLAMOS EN UNA ASEGURADORA UTILIZANDO MACHINE LEARNING,** presentado por: **Marvin Wilfredo Ajcuc Cuzco ,** 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, abril de 2022

AACE/gaoc

ACTO QUE DEDICO A:

Mi esposa	Por su amor y apoyo incondicional en todo momento.
Mi hija	Por ser mi motivación para seguir adelante todos los días.
Mi padre	Por darme su apoyo para iniciar mi carrera.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por ser el lugar que me dio la oportunidad de crecer en mi carrera profesional.
Facultad de Ingeniería	Por brindarme la oportunidad de aprendizaje superior.
Mi asesor	Ing. Denis Fernando Palma Ramírez, por su apoyo, tiempo y disposición para asesorarme.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

GLOSARIO

Asegurado	Persona individual o jurídica que contrata un seguro.
Asegurador	Entidad establecida legalmente que está obligada a responder económicamente por los acuerdos pactados.
Póliza de seguro	El contrato se perfecciona y aprueba mediante documentos privados que se extenderán por duplicado y en este se harán constar los elementos indispensables.
Prima de seguro	Monto económico que debe pagar el asegurado para recibir la obligación pactada.

RESUMEN

La investigación tiene como propósito el diseño de una herramienta tecnológica para el análisis de documentos de reclamo de una aseguradora usando *machine learning,* con el objetivo de disminuir los recursos utilizados actualmente en este proceso y aumentar la calidad de los servicios a través de reducción de tiempos de respuesta y aumento de la confiabilidad, lo que representa una importante ventaja competitiva.

Se detalla el análisis de la situación actual con el objetivo de documentar el estado inicial del proceso, ya que en la parte final del estudio se realizará la comparación cuando la herramienta se encuentre en funcionamiento.

Se procede a evaluar requerimientos de hardware, software y seleccionar servicios para *machine learning.* El desarrollo de la solución tecnológica se da a través del diseño del modelo *machine learning,* el reconocimiento de tipos de documentos, interfaces gráficas para: configuración de documento, configuración de campo y para carga de archivos. Se procede a integrar el sistema OCR y el servicio de *machine learning,* se realiza el entrenamiento de la herramienta, se crean los reportes y se comienza con la etapa de experimentación que se compone de pruebas de confiabilidad y pruebas de capacidad (volumen de procesamiento de datos).

En la presentación de resultados se compara la situación actual versus la implantación de la herramienta tecnológica para los siguientes factores: la confiabilidad, volumen de procesamiento y costos asociados.

1. INTRODUCCIÓN

Muchas empresas no cuentan con alguna herramienta tecnológica que pueda realizar el análisis de documentos y extraer la información de estos. Esto puede deberse a que existe intercambio de información a través de documentos con distintas empresas por lo cual se tienen distintos tipos o diversidad de documentos, es decir no un documento estándar con la información.

En muchos casos no se cuenta con herramientas debido a los costos muy altos en tecnología para poder implementar estas soluciones, así como también puede existir resistencia de algunas áreas en las empresas debido a la implementación de estas herramientas tecnológicas por la idea o temor que sustituya ciertos empleos que intervienen en el proceso de análisis de documentos.

Lo anterior provoca que las empresas puedan perder oportunidades de negocio debido a que el procesamiento de documentos es lento, también los costos pueden elevarse puesto que se debe pagar a personal que se encarga de estas tareas y a consecuencia de esto, la capacidad de procesamiento de los documentos se ve limitado a el personal que se tenga contratado. También, debido a que en el proceso intervienen personas, pueden ocurrir errores humanos debido a confusiones, cansancio, entre otros.

Todo lo anterior podría resolverse contando con una herramienta que sea capaz de procesar los distintos tipos de documentos y que pueda extraer la información correctamente sin importar que se tenga una gran diversidad de documentos. Con esto se podría aumentar la capacidad y reducir el tiempo que

toma el análisis y extracción de la información de estos documentos, por lo cual la empresa tendría una ventaja competitiva.

2. ANTECEDENTES

El análisis, la extracción de información y la clasificación de documentos de forma automatizada es un tema muy complejo y que despierta mucho interés por las aplicaciones y beneficios que puede obtenerse al automatizarse. Todo esto tiene sus bases en el reconocimiento óptico de caracteres (OCR por sus siglas en inglés).

Basarkar (2017) refiere que tuvo como objetivo de investigación: comparar las funciones Bynary, Count y Tfldf y su impacto en la clasificación de documentos. En esta se comparan distintos tipos de vectores a través de los cuales se puede representar un documento y clasificarlo. Compara distintas funciones para evaluar que tan bien funcionan utilizando un conjunto de datos de pruebas de grupos de noticias.

Para este diseño experimental se realizaron distintos escenarios como: eliminar los encabezados y pies de página dado que conduce a una mejor clasificación. Eliminación de palabras consideradas vacías como *“the”, “A”, “From”, “To*”, entre otros. Ya que estas palabras dificultan la clasificación. También se consideró la derivación de palabras, para considerar las palabras que se originan de otras como si fueran la palabra raíz misma. La investigación previamente citada proporciona ejemplos de diseños experimentales que pueden ser útiles al objetivo de analizar y extraer información de documentos, por ejemplo, la eliminación de palabras vacías o la derivación de palabras, algo que puede ayudar a entrenar a la herramienta para poder identificar la información escrita de distintas formas por medio *de machine learning.*

Al procesar documentos con un OCR se obtiene un nivel de confiabilidad del resultado, pero también se puede realizar un post procesamiento para intentar corregir errores y aumentar la confiabilidad final. Para este post procesamiento se puede aplicar machine learning con lo cual se puede aumentar la confiabilidad del procesamiento OCR corrigiendo errores comunes y aplicando otras técnicas (Fonseca, 2019).

Posterior al procesamiento con un OCR se necesita determinar a qué tipo de documento pertenece la información extraída, para lo cual pueden aplicarse distintas técnicas entre las cuales está machine learning. Eckert, Montenegro, López y Candia (2019) refieren que estudiaron la eficacia de un modelo de machine learning para extraer información de textos desestructurados relacionados a historial médico de pacientes. En los historiales clínicos actualmente se manejan muchas herramientas informáticas para tener de forma digital la información de los pacientes, pero existe mucha información histórica que se encuentra en documentos físicos escaneados, es en este punto donde surge la necesidad de extraer la información importante que ayude a los médicos a tomar decisiones y realizar diagnósticos basados en el historial clínico completo de los pacientes.

Se han realizado estudios para definir un modelo que utilice machine learning para poder extraer esa información filtrada y que sea importante en el historial clínico de los pacientes, es decir, no una meramente digitalización, sino que el modelo devuelva datos y predicciones basado en el historial clínico analizado. Esto es muy importante ya que ayuda a los médicos a disminuir el tiempo que pasan leyendo y analizando los documentos digitalizados para poder encontrar información que sea relevante para los casos de cada paciente. Cabe resaltar que en esta investigación el modelo es entrenado utilizando los recursos de IBM Watson. Otro punto importante es que este análisis y extracción de

información se hace sobre documentos escritos en idioma español. La investigación antes citada se relaciona con el presente trabajo en el sentido de la utilización de recursos o servicios que pueden también ser utilizados para hacer el entrenamiento de la herramienta, como IBM Watson. Así como también puede ser muy útil y de gran ayuda la forma en que se aplica al idioma español dado que el caso que analizará en el presente trabajo es sobre reclamos de seguros los cuales están en idioma español.

Otro caso de éxito es el desarrollado por Sainz (2019) en el cual se aplica la inteligencia artificial para el procesamiento y gestión de facturas para una empresa industrial utilizando un robot de software, con el cual logra obtener una mejora significativa en los tiempos y reduce el tiempo que los empleados utilizaban para este procesamiento.

El análisis de los datos es una de las partes más importantes para obtener los mejores y óptimos resultados sin importar a que área pertenezca la información, esta tarea se dificulta cuando la información a procesar es cuantiosa. Ordoñez *et al*. (2019) propusieron y evaluaron una herramienta que realiza búsquedas en documentos del área judicial el cual utiliza inteligencia artificial con lo cual lograron obtener una agilización en la realización de búsquedas y análisis de documentos judiciales. En el ámbito del sistema de justicia de Colombia se utilizan sentencias emitidas previamente por jueces para poder ser utilizadas como bases legales que influyen en la toma de decisiones de casos similares.

Dicho de otra manera, se puede solicitar al sistema de justicia que basen sus decisiones en casos similares que hayan sido resueltos en el pasado. Es en este punto donde se encuentra la dificultad, ya que, debido a la gran cantidad de documentos, la búsqueda basada en alguna base de datos con palabras clave

no es lo suficientemente buena y puede devolver una gran cantidad de resultados no útiles. Aunque existen algunos sistemas ya creados que utilizan inteligencia artificial, estos no funcionan para el idioma español de los documentos. Es por esto por lo que se diseñaron un sistema que genera resúmenes y hace interpretaciones del lenguaje natural para así poder hacer las búsquedas más exactas y eficientes, estos modelos y los algoritmos funcionan utilizando inteligencia artificial. El estudio descrito en el párrafo anterior es muy importante para el presente trabajo ya que tiene un objetivo similar, que es poder procesar y extraer información de documentos que normalmente requieren un procesamiento por parte de humanos, pero aplicando inteligencia artificial se puede realizar de manera más eficiente.

La clasificación, análisis y extracción de la información utilizando OCR ha ido evolucionando continuamente hasta incluir inteligencia artificial para mejorar los resultados (Rishabh, 2020).

Un punto muy importante para tomar en cuenta es la complejidad adicional al procesar escritos a mano a través de un OCR, todos estos escritos son muy importantes en las transacciones humanas y están presentes en un gran porcentaje de documentos de empresas y las técnicas existentes para este fin las cuales se tendrán presentes para incluir el procesamiento de manuscritos en la presente investigación (Jamshed, Maira, Mueen y Rizwan, 2020).

Ahirrao, Baviskar, Kotecha y Vidyasagar (2021) refieren que se han integrado nuevas tecnologías como RPA para poder automatizar el procesamiento de documentos incluyendo el OCR lo cual proporciona un panorama muy bueno del estado actual de las herramientas y técnicas existentes del reconocimiento óptico de caracteres.

En la actualidad se cuentan con distintos motores de OCR los cuales han ido mejorando y evolucionando soportados por grandes empresas y, en algunos casos, por la comunidad ya que son sistemas de código abierto. Estos tienen distintos niveles de confiabilidad, y en algunos casos la confiabilidad depende del idioma a procesar. Heghammer (2021) refiere que encontró que el resultado se ve afectado por el idioma de los documentos a procesar y que Google Document AI obtiene mejores resultados, seguido de Amazon Textract y en tercer lugar Tesseract, los cuales son lideres en motores de OCR.

También en otro ejemplo del uso de inteligencia artificial para procesamiento de documentos es el utilizado por León (2021) en donde utilizó una red neuronal y visión computacional para procesar imágenes y documentos de exámenes en papel para poder así procesar estos documentos y detectar las notas para poder al final tener como resultado un archivo con los datos ya estructurados y con un formato estándar.

3. PLANTEAMIENTO DEL PROBLEMA

3.1. Contexto general

En la actualidad existe un gran intercambio de información entre distintas empresas en su funcionamiento, este intercambio de información muchas veces se realiza a través de documentos lo cuales no contienen la información de forma estructurada o estandarizada por lo cual se genera una gran diversidad de documentos.

En el caso específico de este estudio se analiza el caso de una aseguradora que recibe muchos documentos de reclamos de seguros los cuales debe analizar, este modo de funcionamiento ha sido parte del proceso desde hace varias décadas atrás y se ha intentado resolver mediante la contratación de personal para digitalización y clasificación, la implementación de un sistema donde el personal de digitalización almacena la información procesada. Pero a medida que la aseguradora crece, también crece el volumen de documentos a procesar, lo cual vuelve cada vez más difícil, costoso y tardado el procesamiento de los documentos.

3.2. Descripción del problema

La gran variedad de tipos de documentos por medio de los cuales la empresa intercambia información con otras empresas, proveedores y clientes, provoca que el procesamiento de estos sea lento y costoso, lo cual a su vez provoca que la empresa pueda perder oportunidades de negocio. También los costos pueden elevarse puesto que se debe pagar a personal que se encarga de

estas tareas y a consecuencia de esto, la capacidad de procesamiento de los documentos se ve limitado el personal que se tenga contratado. Además, debido a que en el proceso intervienen personas, pueden ocurrir errores humanos debido a confusiones, cansancio, entre otros. Lo cual se ve reflejado en la confiabilidad del procesamiento de la información.

3.3. Formulación del problema

A continuación, se muestran las preguntas que permitirán la formulación del problema.

3.3.1. Pregunta central

¿Cómo disminuir los tiempos y costos del procesamiento de documentos de reclamos en una aseguradora?

3.3.2. Preguntas auxiliares

- ¿Cómo disminuir el tiempo de procesamiento de la información contenida en documentos?
- ¿Cuál es el porcentaje de confiabilidad que puede obtenerse con una herramienta tecnológica?
- ¿Cuáles son los costos asociados al procesamiento de documentos utilizando machine learning?
- ¿Cómo identificar y clasificar los distintos tipos de documentos?

4. JUSTIFICACIÓN

La realización del presente trabajo se justifica en la línea de investigación de Data Analytics de la Maestría en Ingeniería para la Industria con Especialidad en Ciencias de la Computación y en la línea de investigación de Machine Learning de Ingeniería en Ciencias y Sistemas.

El procesamiento y análisis de documentos es uno de los procesos que se realiza en muchas empresas de distintas industrias, especialmente el sector financiero y de seguros se maneja mucho intercambio de información a través de documentos con sus clientes, proveedores y cualquier empresa con las que tengan relación. Estas clasificaciones para poder extraer la información cuando los documentos son muy diversos normalmente la realizan empleados que analizan y determinan que tipo de documento y que información es la que debe extraerse del mismo. Con la presente investigación se dará una solución para optimizar la clasificación y extracción de la información de esta diversidad de documentos lo cual generará una reducción de costos y tiempos del procesamiento de los documentos.

Se diseñará una herramienta tecnológica que utilice machine learning la cual pueda ser capaz de realizar la clasificación y extracción de la información de la diversidad de documentos relacionados a reclamos del área de seguros, de tal forma que este procesamiento de documentos no necesite ser realizado por humanos, si no que, a través de machine learning, se entrene a la herramienta para poder tener una base de los diversos tipos de documentos y que sea capaz de reconocer y clasificar a que tipo o clasificación pertenece un documento que se le ingrese y pueda extraer la información para la cual fue entrenado.

Esta investigación y la herramienta que se diseñará beneficiará a la empresa aseguradora, especialmente a su Departamento de Reclamos, debido a que le proporcionará una herramienta que podrá reducir sus costos en el procesamiento de documentos, así como también reducirá los tiempos de dicho procesamiento, lo cual también le proporcionará una ventaja competitiva en el sector. Aunque la investigación se enfocará en los documentos asociados a reclamos de seguros, en el futuro puede ser una muy buena base para reutilizarse o expandirse a aplicarse a otros tipos de documentos y a otras empresas de distintas áreas o sectores de la industria.

5. OBJETIVOS

5.1. General

Diseñar una herramienta para disminuir los tiempos y costos al procesar y extraer información de documentos de reclamos de una aseguradora utilizando machine learning.

5.2. Específicos

- Disminuir el tiempo de procesamiento de documentos utilizando una herramienta basada en machine learning.

- Determinar el porcentaje de confiabilidad de la obtención de la información de los documentos utilizando machine learning.

- Determinar los costos asociados al uso de la tecnología necesaria para utilizar servicios de machine learning.

6. NECESIDADES POR CUBRIR Y ESQUEMA DE LA SOLUCIÓN

En muchas empresas no existe una herramienta que permita automatizar el procesamiento de documentos, específicamente para documentos de reclamos de seguros. Por lo cual con la presente investigación se pretende diseñar una herramienta que utilizando machine learning pueda procesar documento y permita extraer la información de estos aun cuando sean de distintos orígenes y con distintos formatos.

Con la variedad de documentos, se necesita tener una forma de que la herramienta pueda reconocer los distintos tipos conforme se le proporcione los documentos para que esta los procese. Esto se conseguirá a través de la utilización de machine learning, entrenando algún modelo para que pueda hacer este reconocimiento de los distintos tipos de documentos y sea capaz de determinar el tipo para poder extraer la información correctamente.

Debido a que la empresa siempre debe minimizar costos y para que el proyecto se lleve a cabo con éxito, se necesita optimizar el modelo y los servicios de machine learning de varios proveedores para poder tener los costos mínimos y que estos proporcionen una diferencia significativa comparado con los costos del procesamiento realizado por el personal de la empresa. Para saber el costo actual se tomarán los datos de los sueldos de los empleados que actualmente tienen como tarea el procesamiento de los documentos de reclamos de seguros incluyendo sus prestaciones, y se obtendrá el dato del total de documentos procesados durante el último año para tener un promedio aproximado del costo de procesar un documento. Con la herramienta se considerarán los costos de la

utilización del servicio de machine learning, los costos de transferencia de datos hacia la nube, almacenamiento y el costo del servicio de cómputo donde se ejecute la herramienta finalmente.

La confiabilidad en el resultado del procesamiento de los documentos y la extracción correcta de la información contenida en estos debe ser lo suficientemente alta comparada con la confiabilidad actual. Para esto el modelo de machine learning y la herramienta serán entrenados con muchos documentos de pruebas y se ira midiendo la confiabilidad y corrigiendo la herramienta para que mejore continuamente. Para esto se determinará la confiabilidad actual por medio de estadísticas de errores que se han reportado con el procesamiento por parte de los empleados y en la fase de pruebas de la herramienta se tendrá que realizar una revisión del resultado del procesamiento para poder ir determinando la confiabilidad y realizar las mejoras correspondientes.

También se medirán los tiempos del procesamiento de documentos por la herramienta, para poder tener promedios y otras estadísticas y así poder determinar la capacidad de procesamiento que la herramienta le proporcionará a la empresa comparada con su modalidad actual en la que muchos empleados se encargan de este procesamiento. Para obtener el dato actual se determinará la cantidad de empleados que se dedican al procesamiento de los documentos por la cantidad de horas de su jornada laboral, y la cantidad del total de documentos procesados durante varios días y así obtener un promedio aproximado de documentos por hora.

7. MARCO TEÓRICO

7.1. Base legal de los seguros

En Guatemala existe la Ley de la Actividad Aseguradora emitida por el congreso de la república de Guatemala, decreto número 25-2010, previo a esta estaba vigente una ley aprobada en el año 1960, el ente rector es la Superintendencia de Bancos, la ley se actualizó considerando:

Que el desarrollo económico y social del país requiere de un sistema de seguros confiable, solvente, moderno y competitivo, que mediante la protección de los bienes asegurados contribuya al crecimiento sostenible de la economía nacional, y que de acuerdo con los procesos de apertura de las economías pueda insertarse adecuadamente en los mercados financieros internacionales (Ley de la Actividad Aseguradora, 2010).

.

7.2. Concepto de seguro

Es un contrato físico o digital mediante el cual el asegurado pacta con la entidad aseguradora el pago de una prima de seguro que le aporta un beneficio económico en caso de ocurrir los siniestros acordados dentro de los límites pactados.

Características:

- Bilateral: ya que ambas partes están obligadas a pagar un monto pactado.
- Oneroso: debido ambas partes incurrirán en un gasto.

- Probable: no se sabe con certeza si el evento asegura ocurrirá o no.

7.3. Elementos básicos en el contexto de seguros

- El asegurador
- El solicitante
- El objeto del seguro
- El riesgo asegurable
- El monto asegurado o el límite de responsabilidad del asegurador, según el caso
- La prima o precio del seguro
- La obligación del asegurador de efectuar el pago del seguro en todo o en parte, según el tamaño del siniestro.

7.3.1. Descripción

- Asegurador: entidad establecida legalmente que está obligada a responder económicamente por los acuerdos pactados.
- Asegurado: persona individual o jurídica que contrata un seguro.
- Prima de seguro: es el monto económico que debe pagar el asegurado para recibir la obligación pactada.
- La póliza: el contrato se perfecciona y aprueba mediante documentos privados que se extenderán por duplicado y en este se harán constar los elementos indispensables. Toda póliza debe contener los siguientes datos:

    - El nombre y domicilio del asegurador
    - Los nombres y domicilios del solicitante, asegurado y beneficiario
    - La calidad en que actúa el solicitante del seguro

- La identificación precisa de la persona o caso con respecto a la cual se contrata el seguro
- La vigencia del contrato con detalle fecha y hora
- El monto asegurado
- La prima
- La naturaleza de los riesgos tomados
- La fecha en que se celebra el contrato
- La firma de las partes
- Todas las cláusulas pactadas

7.3.1.1. Tipos de seguros

Los seguros se clasifican por:

- Duración
- Naturaleza del riesgo
- Personales
- Responsabilidad civil
- Combinados
- Por número de asegurados
- Individuales
- Grupales/colectivos
- Por la clase de asegurador
- Privados
- Obligatorio

7.4. Ramos de seguros en el contexto guatemalteco

En la presente investigación se tratarán los tres ramos principales utilizados en la aseguradora de estudio, pero en la ley de seguros se contempla únicamente dos ramos de vida y de daños.

7.4.1. Daños (vehículos, inmuebles, mercaderías, aviación)

El ramo de daños se ocupa de proteger los bienes tangibles de los siniestros pactados dentro del seguro contratado.

7.4.2. Gastos médicos

Es un contrato de protección financiera en caso de necesitar atención médica por enfermedad o accidente, existen de diferentes coberturas y tienen un límite de acuerdo con la cuota pactada.

7.4.3. Seguros de vida

Es un contrato de protección financiera a la familia o beneficiarios del seguro en caso de muerte de la persona asegurada existen diferentes seguros de vida y diferentes coberturas.

7.5. Reclamos de seguros

Los reclamos de seguro se hacen efectivos cuando ocurren los siniestros que se contemplaron y se pactaron en el contrato de seguro, es decir el reclamo es el pago del monto pactado.

Para hacer el reclamo de seguros hay diferentes procedimientos dependiendo de la entidad que se trate, para hacerlo efectivo la entidad emite un formulario con los principales datos de la prima de seguro.

7.6. Herramientas tecnológicas por considerar

A continuación, se presentan conceptos importantes para entender mejor el presente trabajo de graduación.

7.6.1. Inteligencia artificial

Kaplan y Haenlein (2019) definen la inteligencia artificial como la capacidad de un sistema para interpretar correctamente datos externos, para aprender de dichos datos y emplear esos conocimientos para lograr tareas y metas concretas a través de la adaptación flexible.

La inteligencia artificial es una ciencia joven que se aplica cada vez más a los procesos comerciales, actualmente ya no se relaciona únicamente con algoritmos sino que se puede componer de un conjunto de sistemas tecnológicos que son capaces de replicar las tareas humanas, según Takes (2007) la IA es una rama de las ciencias computacionales encargada de estudiar modelos de cómputo capaces de realizar actividades propias de los seres humanos con base en dos de sus características primordiales: el razonamiento y la conducta.

Algunos ejemplos de inteligencia artificial son: sistemas de conducción autónomos y juegos de ajedrez.

7.6.2. Reconocimiento óptico de caracteres (OCR)

Es un software con la capacidad de convertir una imagen en caracteres que se pueden digitalizar, aunque ya se ha utilizado para diferentes aplicaciones como digitalizar textos, la tarea se dificulta cuando se trata de manuscritos o imágenes con poco contraste entre la letra y el fondo.

7.6.3. Tesseract

Es un motor de OCR basado en una red neuronal, funciona para varios sistemas operativos, es un software libre y su desarrollo es financiado por Google desde el 2006.

Tesseract se utiliza para la detección de texto en dispositivos móviles, en video y en la detección de spam de imágenes de Gmail.

7.6.4. Cloud computing

En español se traduce como la computación de la nube y es una tecnología que ofrece diversos servicios de manera remota, se creó a finales de los 90 ofreciendo principalmente servicios de almacenamiento pero en la actualidad ofrece distintos servicios y herramientas como: redes, servidores y distintos microservicios. La computación de la nuble ha permitido acceder fácilmente a recursos que no se hubiera imaginado en el pasado lo que ha permitido facilitar el desarrollo tecnológico, actualmente las principales empresas que prestan este servicio son: Amazon, Google y Microsoft.

7.6.5. Machine learning

Es parte de la ciencia de la computación y una rama de la inteligencia artificial, consiste en un algoritmo capaz de procesar, analizar y estructurar la información de manera que pueda realizar predicciones y aprender de los análisis realizados para aplicar los conocimientos obtenidos y con base en la experiencia, aplicar el aprendizaje obtenido en acciones futuras sin necesidad de ser reprogramado. Se basa principalmente en modelos matemáticos utilizando para el análisis de datos la inferencia estadística.

Actualmente se utiliza machine learning en diferentes ramas como:

- Medicina
- Economía
- Finanzas
- *Marketing*
- Robótica
- Reconocimiento de lenguaje escrito

Existen principalmente 3 modelos utilizados para machine learning estos son: geométricos, probabilísticos y lógicos. Para cada uno de estos se puede aplicar distintos tipos de algoritmos los cuales pueden ser supervisados o no, así como muchas otras características.

7.6.6. Amazon Web Service (AWS)

Es una empresa que brinda servicios en la nube como: potencia de cómputo, almacenamiento para bases de datos, entrega de contenido, herramientas para crear, entrenar e implementar modelos machine learning,

entre otras, prometiendo mayor flexibilidad, escalabilidad y fiabilidad. Las herramientas se pueden encontrar por categoría de tecnología (Análisis y lagos de datos, machine learning, entre otros) o por sector (publicidad, servicios financieros, tecnología para videojuegos, entre otros).

- AWS SageMaker

Es un servicio de aprendizaje automático, que permite a científicos de datos y desarrolladores a preparar, crear, entrenar e implementar con rapidez modelos de aprendizaje automático de alta calidad al poner a disposición un amplio conjunto de capacidades especialmente creadas para el aprendizaje automático. Amazon lo promociona ofreciendo ventajas de aceleración de innovación con herramientas especialmente creadas para cada paso del desarrollo de aprendizaje automático, incluidos la aplicación de etiquetas, la preparación de los datos, la ingeniería de características, la detección de tendencias estadísticas, el aprendizaje automático automatizado, el entrenamiento, la adaptación, el alojamiento, el monitoreo y los flujos de trabajo. Empresas reconocidas utilizan este servicio, Amazon indica que es el servicio con más rápido crecimiento hasta la fecha.

- Amazon Textract

Amazon Textract es un servicio de machine learning que extrae texto escritura a mano y datos de documentos escaneados de forma automática. Va más allá del simple reconocimiento óptico de caracteres (OCR) ya que es capaz de indentificar, comprender y extraer datos de formularios y tablas. En la actualidad, muchas empresas extraen datos de documentos como archivos PDF, imágenes, tablas y formularios escaneados de forma manual o mediante un software de OCR simple que requiere una configuración manual y a menudo

exige una reconfiguración cuando cambia de formulario. Textract, para superar estos procesos manuales y costosos, utiliza el machine learning a fin de leer y procesar cualquier tipo de documento y extraer con precisión texto, escritura a mano, tablas y otros datos sin esfuerzo manual. Puede automatizar el procesamiento de documentos y tomar medidas sobre la información que se extrae, ya sea mediante la automatización del procesamiento de préstamos o extracción de información de facturas y recibos. Textract puede extraer los datos en minutos en lugar de horas o días. Además, puede agregar revisiones humanas con Amazon Augmented AI para supervisar los modelos y llevar a cabo revisiones de información confidencial.

Casos de uso:

- Servicios financieros
- Sector sanitario y de ciencias biológicas
- Sector público

7.6.7. Google Cloud Platform (GCP)

Es una empresa dedicada a ofrecer servicios en la nube a través de una plataforma en la cual ofrecen más de 90 servicios de tecnología de la información, facilitando herramientas que están listas para ser usadas, por ejemplo: ejecución de máquina rápidamente gracias a sus configuraciones predefinidas, crea máquinas virtuales con la cantidad óptima de vCPUs y memoria, abarata la computación a través de máquinas interrumpibles, encriptación de datos, entre otros. Algunas de las empresas que utilizan Google cloud son: Paypal, The home Depot, Spotify, Twitter, entre otras.

- Google datalab

Cloud Datalab es un entorno interactivo de análisis de datos y aprendizaje automático diseñado para Google Cloud Platform. Puede usarse a fin de explorar, analizar, transformar y visualizar datos de forma interactiva, y para generar modelos de aprendizaje automático a partir de datos mediante lenguajes como Python y SQL. La plataforma ofrece instructivos y ejemplos de algunas tareas que se pueden realizar para familiarizarse con las distintas herramientas, incluye un conjunto de bibliotecas de Phython de código abierto, también agrega bibliotecas para acceder a servicios clave de Google Cloud Platform como: Google BigQuery, Google Machine Learning Engine, Google Dataflow y Google Cloud Storage.

- Google Document AI Solution

Automatiza la captura de datos a escala para reducir los costes de procesamiento de documentos. Extrae datos estructurados de documentos sin estructurar y los pone a disposición de las aplicaciones y los usuarios para mejorar la eficiencia operativa. Automatiza y valida todos los documentos para optimizar los flujos de trabajo de cumplimiento, reduciendo dificultades y logrando que los datos sean precisos y se ajusten a las normativas en todo momento. Toma decisiones más acertadas basándose en los datos de los documentos y colocándolos también a disposición de las aplicaciones y los usuarios.

Automatiza y valida los datos para que los flujos de trabajo sean más eficientes y no supongan ninguna dificultad. Ofrece altos estándares de seguridad para la protección de datos. Aprovecha los datos de los documentos para obtener información nueva y valiosa sobre los productos y satisfacer las expectativas de los clientes.

7.6.8. Microsoft Azure

Es una plataforma de Microsoft que ofrece servicios en la nube está compuesta por más de 200 productos y servicios diseñados para brindar soluciones que permitan resolver dificultades actuales. Permite crear, ejecutar y administrar aplicaciones en varias nubes, en el entorno local y en el perímetro, con las herramientas y los marcos que se prefiera. Ofrece altos niveles de confianza, transparencia y cumplimiento de estándares y normativos, según datos de la página de Microsoft el 95 % de las empresas de la lista de Fortune 500 confía en Azure para obtener servicios en la nube, brindado servicios a empresas de distintos tamaños y niveles de experiencia. Es compatible con tecnologías de código abierto, por lo que se puede utilizar con las herramientas y tecnologías que se prefiera.

- Azure computer visión

Es un servicio de inteligencia artificial que analiza contenido en imágenes y video, automatiza la extracción de texto, utiliza el procesamiento de datos visuales para etiquetar el contenido con objetos, conceptos y extracción de texto, genera descripciones de imágenes, se ha utilizado para la automatización de proceso robóticos, la gestión de activos digitales y el análisis espacial; que permite comprender cómo se mueven las personas en un espacio físico ya sea una oficina o una tienda, permite contar personas en una habitación, rastrear rutas, comprender los tiempos de permanencia frente a una estantería y determinar los tiempos de espera en las colas. Se ha utilizado en diversas áreas como:

  - Economía
  - *Marketing*
  - Logística

- Ventas
- Medicina

Computer Vision garantiza un disponibilidad del 99.9 % como se encuentra en el sitio oficial de Google Cloud.

- Azure machine learning studio

Es un servicio independiente y modernizado que ofrece una plataforma de ciencia de datos completa. Admite las experiencias de los tipos código primero y código bajo. Es un portal web para desarrolladores y científicos de datos de Azure Machine Learning combina las experiencias de los tipos sin código y código primero para una plataforma de ciencia de datos inclusiva. En la página ofrecen manuales de uso y recomendaciones para su uso como: la creación de proyectos de aprendizaje automático, como administrar activos y recursos de aprendizaje automático directamente en el explorador, Studio puede simplificar el modo en que se administran los recursos del área de trabajo incluso para desarrolladores experimentados, algunas de sus características:

- Clústeres de proceso escalables para el aprendizaje a gran escala.
- Seguridad y gobernanza para empresas.
- Interoperabilidad con las herramientas de código abierto conocidas.
- MLOps integral.

Permite control de versiones de entidades (modelos, datos, flujos de trabajo), automatización de flujos de trabajo, integración con herramientas de CICD, implementaciones de CPU y GPU entre otras.

8. PROPUESTA DE ÍNDICE DE CONTENIDOS

9. METODOLOGÍA

9.1. Tipo de estudio

El tipo de estudio es cuantitativo debido a que se diseñará una herramienta tecnológica para reducir los tiempos, costos del procesamiento y extracción de información de documentos de reclamos de seguros. Posteriormente se realizarán mediciones para determinar el cumplimiento de los objetivos.

9.2. Diseño

El diseño de la propuesta es descriptivo no experimental debido a que se analizará el estado y la situación actual del funcionamiento del proceso de clasificación y extracción de información de documentos, para luego comparar el proceso con la herramienta tecnológica propuesta.

9.3. Alcance

El alcance será comparativo, se tomará parámetros de medición antes y después de la implantación de la herramienta con el objetivo de determinar la mejora obtenida para la extracción de información de documentos de reclamos.

9.4. Variables e indicadores

Las variables por utilizar en el estudio se describen en la siguiente tabla:

Tabla I. **Variables**

Variables	Definición	Indicadores
Tiempo de procesamiento de documentos	El tiempo necesario para realizar el procesamiento de un documento de reclamo.	Tiempo de procesamiento de reclamo.
Costos asociados al procesamiento de documentos.	Son todos los costos que conlleva el procesamiento de documentos de reclamo.	Costo por procesamiento de reclamo.
Volumen de documentos.	Se refiere al volumen de documentos de reclamo que se procesan.	Documentos procesados al día
Confiabilidad en la clasificación de la información.	Mide la confiabilidad de la información al ser extraída de los documentos de reclamo.	Confiabilidad de Extracción de datos.

Fuente: elaboración propia.

9.5. Fases del estudio

Las fases del estudio aplicadas a la investigación se describen a continuación.

9.5.1. Revisión documental

En esta fase se obtendrá todo el conocimiento teórico concerniente a la herramienta tecnológica que se desea desarrollar para dar solución al problema expuesto. Se investigará en distintas fuentes de información como los son: Libros, artículos científicos y tesis.

La presente investigación trata acerca de la creación de una herramienta tecnológica de extracción de información de documentos de reclamo en una aseguradora con el objetivo de reducir los tiempos y costos del procesamiento de datos, para lo cual será necesario establecer ciertas bases de conocimiento mediante la investigación de los siguientes temas:

- Inteligencia artificial
- OCR
- Machine learning
- Cloud computing

9.5.2. Diseño de instrumentos de recolección de información

Esta fase consiste en el análisis y creación de los formatos que se utilizarán para obtener la información que se considere necesaria para realizar las distintas mediciones a realizar en cada fase del proyecto.

- Formato de observación de procesamiento de reclamos

Tabla II. **Formato de observación**

Formato de observación **Fecha:_____ No. de observación_____**			
No.	**Operación**	**Descripción**	**Tiempo requerido**
			(en segundos)
			##
			##
			##
			##
		Tiempo total del proceso	**##**

Fuente: elaboración propia

- Formato de recursos utilizados para el procesamiento de datos

Tabla III. **Formato de recursos**

No.	**Descripción del factor**	**Costo asociado**
1		
2		
3		
	Total	##########

Fuente: elaboración propia.

- Reporte de volumen de documentos y medición de confiabilidad de documentos de reclamo mediante pruebas.

Tabla IV. **Reporte**

Reporte de procesamiento de reclamos			
Fecha	**Código**	**Tipo de reclamo**	**Concluido/En proceso**

Estadísticos obtenidos	
Promedio de reclamos atendidos al día	###
Varianza	###

Fuente: elaboración propia.

9.5.3. Diseño de la solución tecnológica

En esta fase se procederá a estudiar y diseñar los elementos necesarios para la creación de la herramienta tecnológica, los mismos se describen a continuación:

- Tipos de documentos y campos de información a extraer
    - Se necesita tener un conjunto de documentos que servirán para alimentar el modelo de machine learning y entrenarlo en el reconocimiento de los distintos tipos y la información que cada uno contiene.

- Interfaz gráfica de usuario para configurar los distintos tipos de documentos
- Interfaz gráfica de usuario para configurar los campos de información correspondientes a los distintos tipos de documentos
- Interfaz gráfica de usuario para cargar archivos a la herramienta e iniciar el procesamiento
- Modulo para realizar el entrenamiento de la herramienta con los distintos tipos de documentos y sus campos de información
- Reporte con el resultado de la clasificación y extracción de la información de los documentos
- Requerimientos de software y hardware:
    - Hardware como scanners para obtener distintos tipos de documentos que no se encuentran digitalizados.
    - Software y servicios proporcionados por la nube seleccionada para trabajar los modelos de machine learning.

Hardware de computadoras y dispositivos móviles para utilizar los sistemas e interfaces de usuario.

9.5.4. Desarrollo de la solución

En esta fase se llevará a cabo el desarrollo de la herramienta tecnológica que dará solución al problema expuesto usando como base el diseño previamente realizado.

Para tener un desarrollo ágil y dinámico de la herramienta se utilizará una metodología ágil como Scrum para poder adaptarse conforme se avanza en la investigación. Se tendrán objetivos pequeños y alcanzables para poder obtener al final de la investigación una solución funcional.

Se realizarán las siguientes actividades:

- Investigación de distintos tipos de OCR disponibles para evaluar y elegir el más apto y con mejores resultados realizando pruebas con distintos documentos.
- Investigación de los distintos servicios para machine learning de los distintos proveedores existentes para elegir el que mejor se adapte con los resultados y el presupuesto.
- Diseño del modelo de machine learning.
- Integración del sistema OCR y el servicio de machine learning.
- Entrenamiento del modelo de machine learning alimentando con distintos tipos de documentos y definiendo los campos de información de cada documento.
- Pruebas de ejecuciones sobre documentos para evaluar los resultados obtenidos con la herramienta.
- Dependiendo de los resultados obtenidos se volverá a realizar un entrenamiento para mejorar la confiabilidad.
- Determinación del porcentaje de confiabilidad mínimo que puede otorgar la herramienta para clasificar los documentos.
- Determinación del porcentaje de confiabilidad mínimo que puede otorgar la herramienta para la obtención de los campos de información de los distintos tipos de documentos.
- Pruebas de integración para la interfaz de usuario.

9.5.5. Experimentación

Al finalizar el desarrollo de la herramienta realizará la revisión de la misma para verificar el funcionamiento correcto mediante tres tipos de pruebas que se describen a continuación:

- Pruebas de confiabilidad de clasificación del tipo de documento
- Pruebas de confiabilidad de extracción de información de los documentos
- Pruebas de tiempo de procesamiento de documentos

Se deberá dejar evidencia de las pruebas que se realicen a la herramienta para lo cual se utilizará uno de los formatos anteriormente descrito.

9.5.6. Recolección y evaluación de resultados

En esta fase se realizará un análisis cuantitativo con los datos obtenidos en cada fase del proyecto y se procederá a evaluar los resultados obtenidos.

- Resultados en la pruebas realizadas a la herramienta

Se obtendrá del reporte que se genere de la herramienta que obtendrá información para obtener los siguientes indicadores:

  - Confiabilidad de tipo de documento
  - Confiabilidad de extracción de información
  - Tiempo de procesamiento de documento

- Tiempos de procesamientos de reclamos sin la herramienta tecnológica

Se realizará la medición del sistema utilizado actualmente mediante el uso del formato de observación de procesamiento de reclamos.

- Resultados de los costos en el procesamiento de reclamos

  - Fase inicial: se utilizará el formato de recursos utilizados para el procesamiento de datos, con el objetivo de establecer los costos que conlleva actualmente el sistema para procesar reclamos.

- Fase final: se utilizará el formato de recursos utilizados para el procesamiento de datos, con el objetivo de establecer los costos utilizando la herramienta y poder calcular la diferencia en costos.

9.5.7. Redacción de informe final

En esta fase se procederá a elaborar el informe final con los resultados obtenidos en cada fase del proyecto y utilizando como base el índice propuesto.

9.6. Técnicas de recolección de información

Esta fase describe las técnicas de recolección de datos que se utilizará en la presente investigación, las cuales se detallan a continuación:

9.6.1. Observación de campo

Es necesario comprender la forma en que se procesan los reclamos actualmente para lo cual será necesario realizar el estudio del proceso mediante su observación. Será necesario crear un diagrama de proceso para establecer los tiempos asociados a este.

9.6.2. Observación directa

Se utilizará esta técnica para determinar el diagrama de procesos y también al utilizar los formatos previamente descritos. Para esto será necesario observar el sistema actual por cinco días en un periodo de 2 horas al día.

9.6.3. Recolección de la información

La recolección de datos se obtendrá de dos fuentes la cuales son:

- Datos recolectados mediante formatos
- Datos recolectados mediante indicadores de la herramienta.

10. TÉCNICAS DE ANÁLISIS DE INFORMACIÓN

La información se analizará por medio de estadística descriptiva para calcular los parámetros básicos de los datos, mismos que se utilizarán para validar los objetivos planteados. A continuación, se explicará la manera en que se obtendrá y analizará la información.

10.1. Variables estadísticas

Se utilizarán variables cuantitativas ya que se realizará la medición de la mejora que se obtenga con la utilización de la herramienta tecnológica.

10.1.1. Variables cuantitativas

- Tiempo de procesamiento de reclamo
- Costos por procesamiento de reclamo
- Documentos procesados al día
- Confiabilidad de extracción de datos

10.1.2. Evaluación inicial

Se realizará una medición para determinar las variables cuantitativas que la empresa maneja actualmente sin el uso de una herramienta tecnológica con el objetivo de tener un parámetro de medición y posteriormente verificar la mejora al implementar la herramienta tecnológica.

10.1.3. Evaluación final

Los datos se obtendrán del reporte que genere la herramienta tecnológica el cual tendrá la siguiente información:

- Tiempo de procesamiento de 1 reclamo
- Documentos procesados al día
- Confiabilidad de extracción de datos
    - Confiabilidad de tipo de documento
    - Confiabilidad de extracción de información

10.2. Obtención de datos

Establecida la información de interés se procederá a describir la manera en que se obtendrá los datos en su etapa inicial (sin el uso de la herramienta tecnológica) y en la etapa final (con el uso de la herramienta tecnológica).

10.2.1. Etapa inicial

- Los tiempos de procesamiento (sin la herramienta tecnológica): se obtendrán a través de observación directa para determinar el tiempo de procesar 1 documento de reclamo apoyándose en el siguiente formato de observación:

Tabla V. **Formato de observación**

Formato de observación			
Fecha:_____			**No. de observación_____**
No.	**Operación**	**Descripción**	**Tiempo requerido (segundos)**
			##
			##
			##
			##
			##
			##
		Tiempo total del proceso	**##**

Fuente: elaboración propia.

- Los costos de procesamiento de reclamos (sin la herramienta tecnológica): para calcularlos se tendrá en cuenta los factores que se enumeran en la siguiente tabla:

Tabla VI. **Costos de procesamiento**

No.	Descripción del factor	Costo asociado
1	Sueldo por hora	##########
2	No. de empleados	##########
3	Re trabajo por errores	##########
	Total	##########

Fuente: elaboración propia.

- Los documentos procesados al día (sin la herramienta tecnológica): se obtendrán de las estadísticas que maneja la empresa actualmente.

- Los datos se guardan en una hoja de Excel y contienen los siguientes datos:

Tabla VII. **Reporte de procesamiento**

Reporte de procesamiento de reclamos			
Fecha	**Código**	**Tipo de reclamo**	**Concluido/En proceso**

Estadísticos obtenidos	
Promedio de reclamos atendidos al día	###
Varianza	###

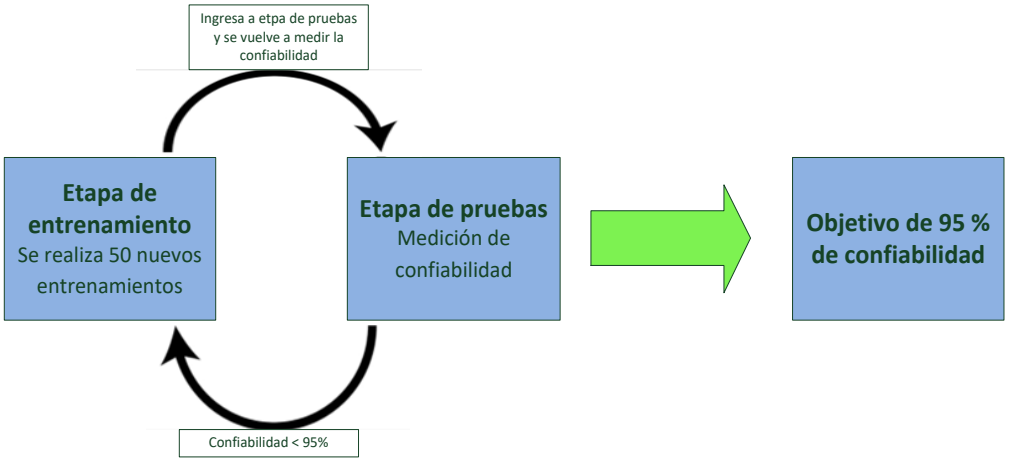
Fuente: elaboración propia.

10.2.2. Etapa final

Los datos en esta etapa se obtendrán a través de las distintas fases de la creación de la herramienta.

- Confiabilidad de tipo de documento y confiabilidad de extracción de información: estos datos se obtendrán en las etapas de entrenamiento y pruebas, la etapa de entrenamiento es cuando se le muestra a la herramienta el tipo de documento y la información que debe extraer de cada tipo de documento, la etapa de pruebas será cuando se ponga en marcha la herramienta y se mida su confiabilidad para determinar el % de confiabilidad en ese momento si la medición es menor al 95 % regresa a la etapa de entrenamiento, se realizará de forma recursiva hasta alcanzar el 95% de confiabilidad, para este caso particular no se puede inferir cuántos entrenamientos necesita la herramienta ya que en cada entrenamiento mejora la confiabilidad pero se desconoce cuánto mejora, el objetivo será ir midiendo cuánto mejora con cada 50 entrenamientos realizados, con estos datos se podrá graficar y verificar la línea de tendencia que presenta. A continuación, se explica gráficamente lo expuesto con anterioridad.

Figura 1. **Diagrama de entrenamiento y pruebas**



Fuente: elaboración propia, usando Visio.

La información obtenida en estas etapas se guardará en una tabla como se muestra a continuación:

Tabla VIII. **Control de entrenamientos**

Tabla de control de entrenamientos		
Etapa de entrenamiento	**Pruebas**	**Medición de confiabilidad**
1	1	30 %
2	1	30 %
3	1	30 %
4	1	30 %
5	1	30 %
6	1	30 %
7	1	30 %
8	1	30 %
.	1	30 %
.	1	30 %
.	1	30 %
.	1	30 %
50	1	30 %
1	2	37 %
.	2	37 %
.	2	37 %
100	2	37 %
.	.	.
.	.	.

Fuente: elaboración propia.

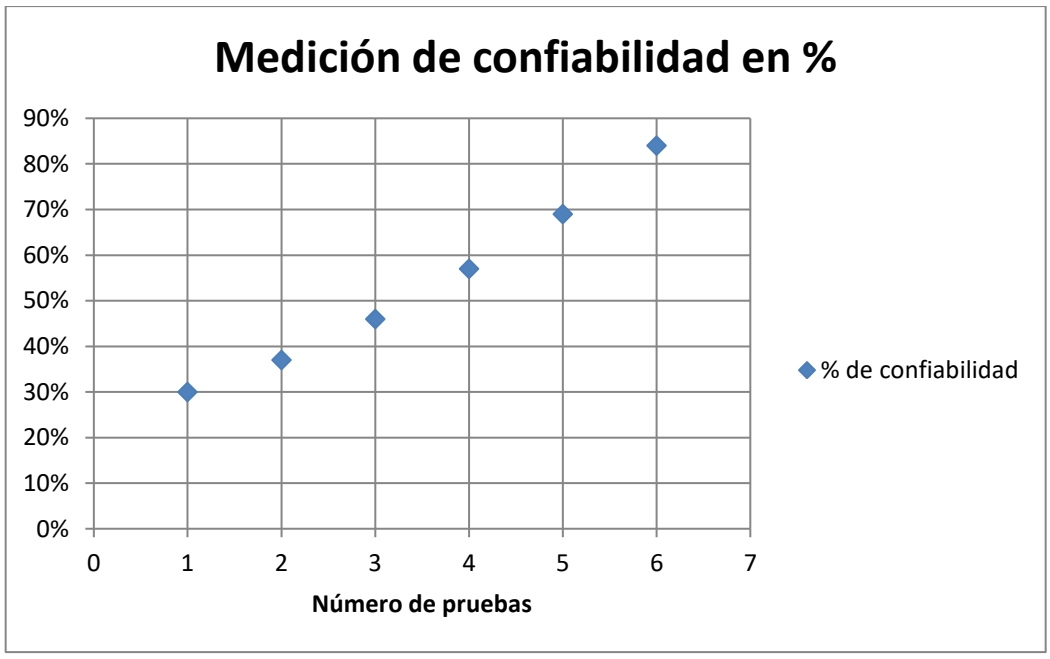
Los datos se procesarán a la siguiente tabla para realizar el gráfico de comportamiento de los datos.

Tabla IX. **Control de pruebas**

Tabla: control de pruebas	
Número de prueba	**Porcentaje de confiabilidad**
1	30 %
2	37 %
3	46 %
4	57 %
5	69 %
6	84 %
.	
.	

Fuente: elaboración propia.

Figura 2. **Gráfico de medición de confiabilidad**



Fuente: elaboración propia.

Realizado el gráfico se procederá a verificar la línea de tendencia del gráfico.

- Tiempo de procesamiento de 1 reclamo: este se obtendrá al poner en marcha la herramienta de la tabla presentada anteriormente agregando una columna de tiempo de procesamiento, para luego obtener la media, determinando que la media= tiempo de procesamiento de 1 reclamo.

Tabla X. **Tabla de control de entrenamientos**

Tabla de control de entrenamientos			
Etapa de entrenamiento	**Pruebas**	**Medición de confiabilidad**	**Tiempo de procesamiento en segundos**
1	1	30 %	2
2	1	30 %	3
3	1	30 %	4
4	1	30 %	.
5	1	30 %	.
6	1	30 %	.
7	1	30 %	.
8	1	30 %	.
.	1	30 %	.
.	1	30 %	.
.	1	30 %	.
.	1	30 %	.
50	1	30 %	.
51	2	37 %	.
.	2	37 %	.
.	2	37 %	.
100	2	37 %	.
.	.	.	.
.	.	.	.

Fuente: elaboración propia.

- Documentos procesados al día

Esta información se obtendrá con la fecha de la prueba los datos se muestran en la siguiente tabla:

Tabla XI. **Documentos procesados al día**

Fecha de la prueba	Pruebas	Medición de confiabilidad	Tiempo de procesamiento en segundos
01/01/2022	1	30 %	2
01/01/2022	1	30 %	3
01/01/2022	1	30 %	4
01/01/2022	1	30 %	.
01/01/2022	1	30 %	.
03/01/2022	1	30 %	.
03/01/2022	1	30 %	.
03/01/2022	1	30 %	.
.	2	37 %	.
.	2	37 %	.
.	2	37 %	.
.	2	37 %	.
.	.	.	.
.	.	.	.

Fuente: elaboración propia.

Se obtendrá los procesamientos realizados al día para establecer este indicador obteniendo la media de pruebas por día.

10.3. Resultados

Calculados los parámetros iniciales y finales se procederá a realizar la medición de la mejora obtenida con la implementación de la herramienta tecnológica y se procederá a interpretar los resultados. El formato para realizar la comparación se presenta a continuación:

10.3.1. Propuesta de comparación

A continuación, se muestra una tabla donde se realiza la comparación de resultados.

Tabla XII. **Comparación de mejora**

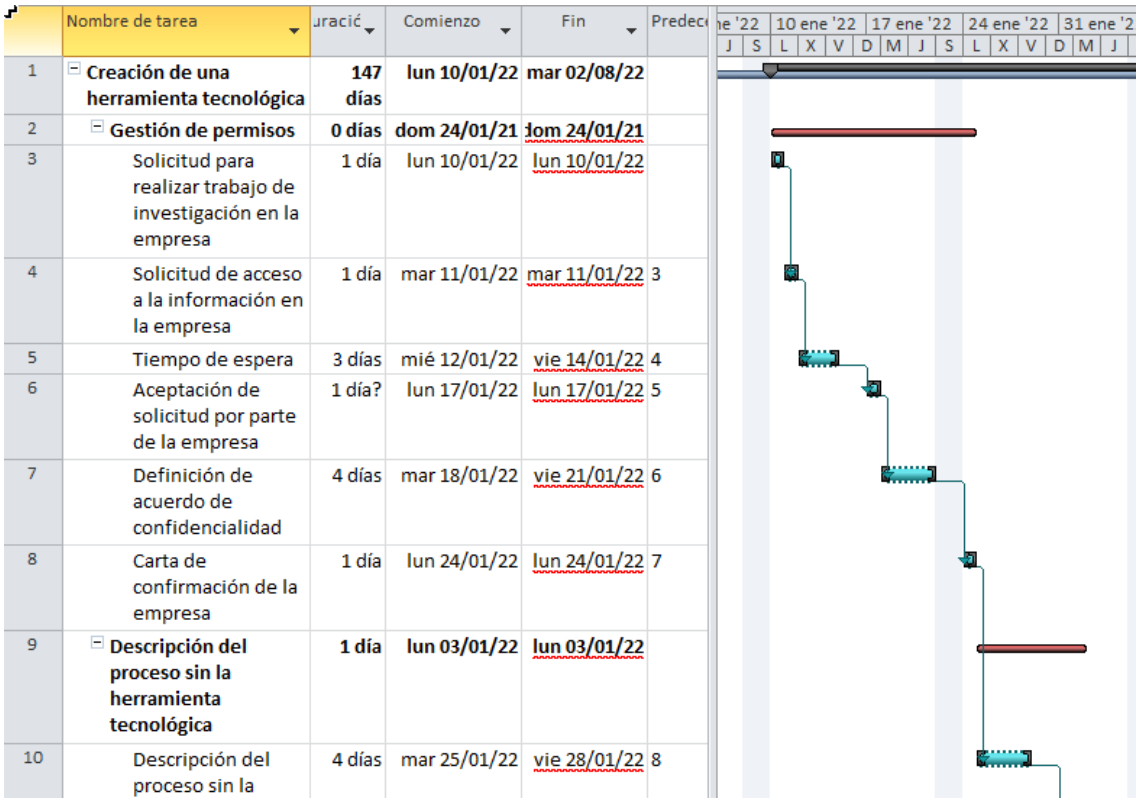
Nombre de la variable	Parámetro inicial	Parámetro final	Porcentaje de mejora
Tiempo de procesamiento de reclamo	a	B	Medición de la mejora
Costo por procesamiento de reclamo	c	D	Medición de la mejora
Documentos procesados al día	e	F	Medición de la mejora
Confiabilidad en la extracción de datos	g	H	Medición de la mejora

Fuente: elaboración propia.

11. CRONOGRAMA

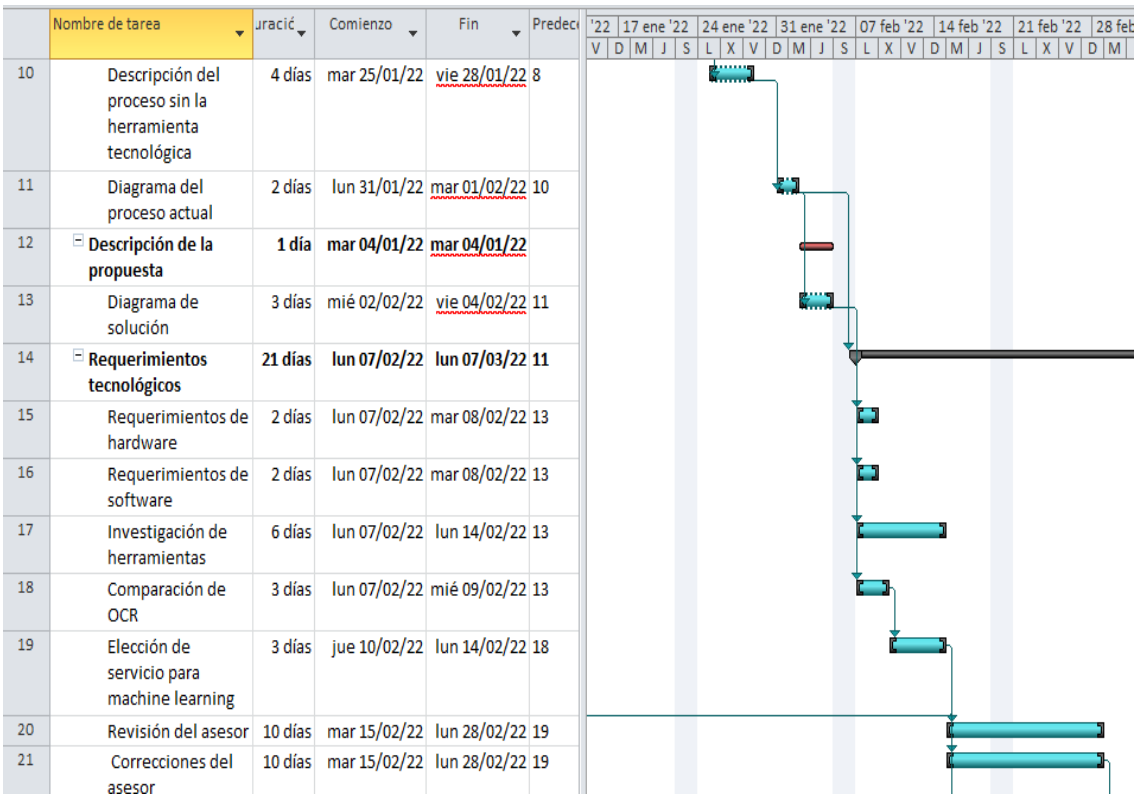
A continuación, se presenta el cronograma del proyecto.

Figura 3. **Lista de tareas y diagrama de Gantt**



	Nombre de tarea	Duración	Comienzo	Fin	Predec
1	**Creación de una herramienta tecnológica**	**147 días**	**lun 10/01/22**	**mar 02/08/22**	
2	**Gestión de permisos**	**0 días**	**dom 24/01/21**	**dom 24/01/21**	
3	Solicitud para realizar trabajo de investigación en la empresa	1 día	lun 10/01/22	lun 10/01/22	
4	Solicitud de acceso a la información en la empresa	1 día	mar 11/01/22	mar 11/01/22	3
5	Tiempo de espera	3 días	mié 12/01/22	vie 14/01/22	4
6	Aceptación de solicitud por parte de la empresa	1 día?	lun 17/01/22	lun 17/01/22	5
7	Definición de acuerdo de confidencialidad	4 días	mar 18/01/22	vie 21/01/22	6
8	Carta de confirmación de la empresa	1 día	lun 24/01/22	lun 24/01/22	7
9	**Descripción del proceso sin la herramienta tecnológica**	**1 día**	**lun 03/01/22**	**lun 03/01/22**	
10	Descripción del proceso sin la	4 días	mar 25/01/22	vie 28/01/22	8

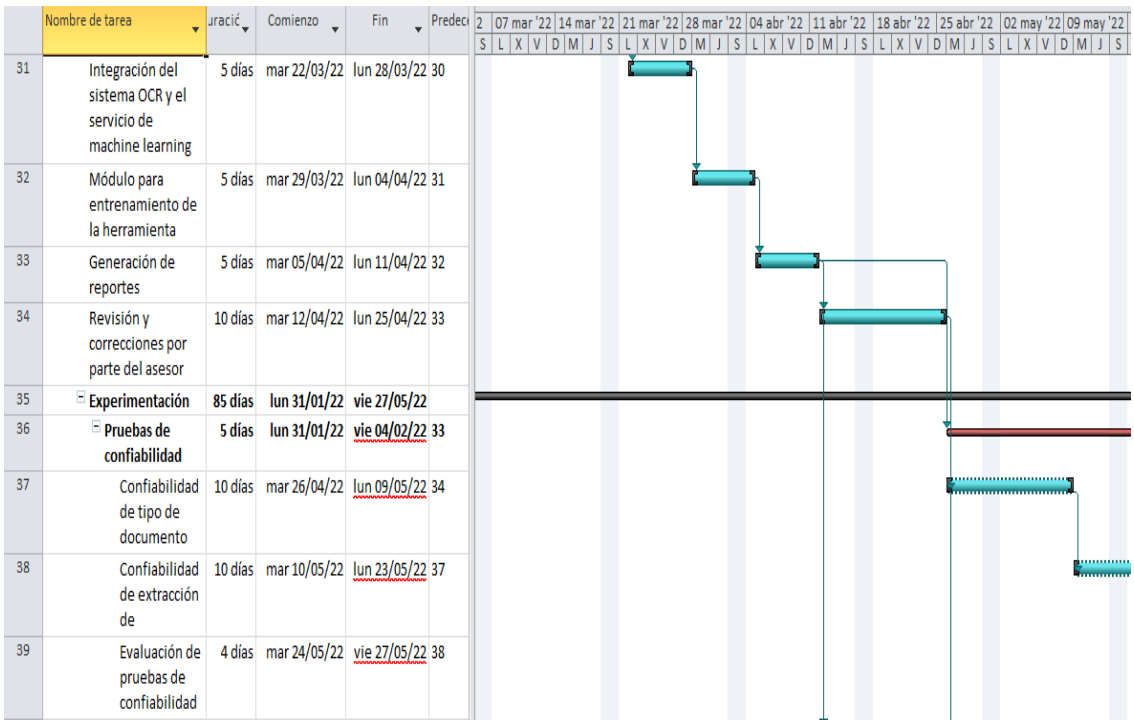
Continuación figura 3.



	Nombre de tarea	Duración	Comienzo	Fin	Predec
10	Descripción del proceso sin la herramienta tecnológica	4 días	mar 25/01/22	vie 28/01/22	8
11	Diagrama del proceso actual	2 días	lun 31/01/22	mar 01/02/22	10
12	**Descripción de la propuesta**	**1 día**	**mar 04/01/22**	**mar 04/01/22**	
13	Diagrama de solución	3 días	mié 02/02/22	vie 04/02/22	11
14	**Requerimientos tecnológicos**	**21 días**	**lun 07/02/22**	**lun 07/03/22**	**11**
15	Requerimientos de hardware	2 días	lun 07/02/22	mar 08/02/22	13
16	Requerimientos de software	2 días	lun 07/02/22	mar 08/02/22	13
17	Investigación de herramientas	6 días	lun 07/02/22	lun 14/02/22	13
18	Comparación de OCR	3 días	lun 07/02/22	mié 09/02/22	13
19	Elección de servicio para machine learning	3 días	jue 10/02/22	lun 14/02/22	18
20	Revisión del asesor	10 días	mar 15/02/22	lun 28/02/22	19
21	Correcciones del asesor	10 días	mar 15/02/22	lun 28/02/22	19

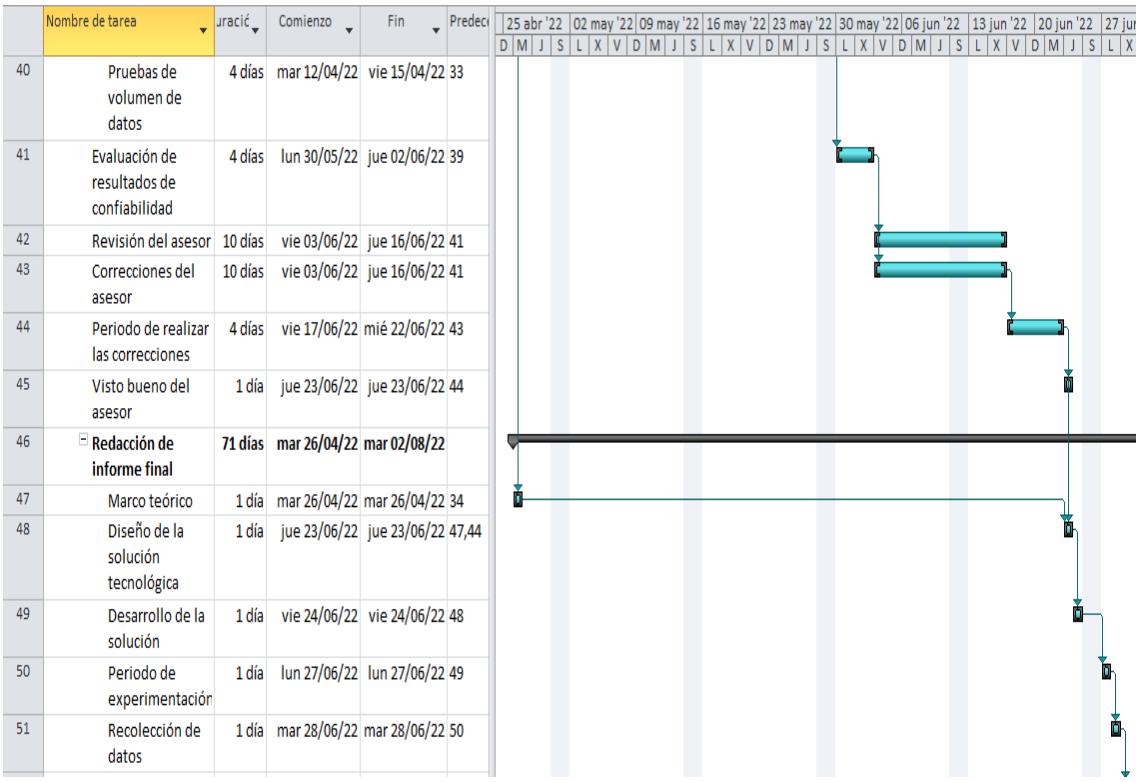
Continuación figura 3.

	Nombre de tarea	Duración	Comienzo	Fin	Predec
22	Periodo de realizar las correcciones	5 días	mar 01/03/22	lun 07/03/22	21
23	Revisión del asesor	1 día	mar 01/03/22	mar 01/03/22	21
24	Visto bueno del asesor	1 día	mié 02/03/22	mié 02/03/22	23
25	**Desarrollo de la solución tecnológica**	**144 días**	**jue 13/01/22**	**mar 02/08/22**	**19**
26	**Diseño del modelo machine learning**	**15 días**	**jue 13/01/22**	**mié 02/02/22**	**19**
27	Reconocimiento de tipo de documento	10 días	mar 15/02/22	lun 28/02/22	19
28	Interfaz gráfica de usuario para configuración de documento	5 días	mar 01/03/22	lun 07/03/22	27
29	Interfa gráfica de usuario para configuración de campos	5 días	mar 08/03/22	lun 14/03/22	28
30	Interfaz grafica de usuario para cargar archivos	5 días	mar 15/03/22	lun 21/03/22	29

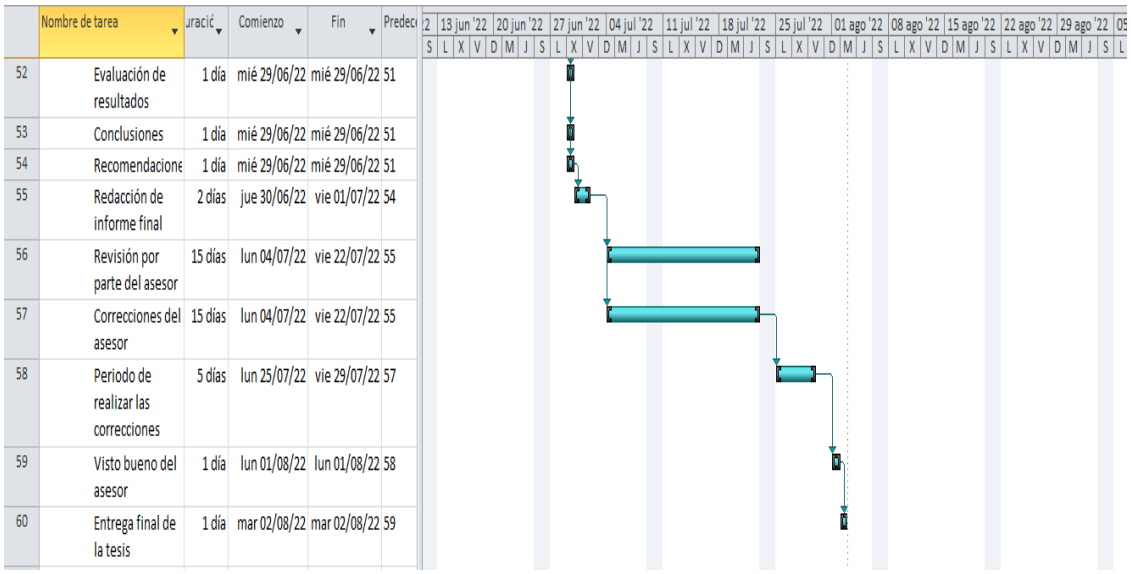


	Nombre de tarea	Duración	Comienzo	Fin	Predec
31	Integración del sistema OCR y el servicio de machine learning	5 días	mar 22/03/22	lun 28/03/22	30
32	Módulo para entrenamiento de la herramienta	5 días	mar 29/03/22	lun 04/04/22	31
33	Generación de reportes	5 días	mar 05/04/22	lun 11/04/22	32
34	Revisión y correcciones por parte del asesor	10 días	mar 12/04/22	lun 25/04/22	33
35	**Experimentación**	**85 días**	**lun 31/01/22**	**vie 27/05/22**	
36	**Pruebas de confiabilidad**	**5 días**	**lun 31/01/22**	**vie 04/02/22**	**33**
37	Confiabilidad de tipo de documento	10 días	mar 26/04/22	lun 09/05/22	34
38	Confiabilidad de extracción de	10 días	mar 10/05/22	lun 23/05/22	37
39	Evaluación de pruebas de confiabilidad	4 días	mar 24/05/22	vie 27/05/22	38

Continuación figura 3.



	Nombre de tarea	Duración	Comienzo	Fin	Predec
40	Pruebas de volumen de datos	4 días	mar 12/04/22	vie 15/04/22	33
41	Evaluación de resultados de confiabilidad	4 días	lun 30/05/22	jue 02/06/22	39
42	Revisión del asesor	10 días	vie 03/06/22	jue 16/06/22	41
43	Correcciones del asesor	10 días	vie 03/06/22	jue 16/06/22	41
44	Periodo de realizar las correcciones	4 días	vie 17/06/22	mié 22/06/22	43
45	Visto bueno del asesor	1 día	jue 23/06/22	jue 23/06/22	44
46	**Redacción de informe final**	**71 días**	**mar 26/04/22**	**mar 02/08/22**	
47	Marco teórico	1 día	mar 26/04/22	mar 26/04/22	34
48	Diseño de la solución tecnológica	1 día	jue 23/06/22	jue 23/06/22	47,44
49	Desarrollo de la solución	1 día	vie 24/06/22	vie 24/06/22	48
50	Periodo de experimentación	1 día	lun 27/06/22	lun 27/06/22	49
51	Recolección de datos	1 día	mar 28/06/22	mar 28/06/22	50



	Nombre de tarea	Duración	Comienzo	Fin	Predec
52	Evaluación de resultados	1 día	mié 29/06/22	mié 29/06/22	51
53	Conclusiones	1 día	mié 29/06/22	mié 29/06/22	51
54	Recomendacione	1 día	mié 29/06/22	mié 29/06/22	51
55	Redacción de informe final	2 días	jue 30/06/22	vie 01/07/22	54
56	Revisión por parte del asesor	15 días	lun 04/07/22	vie 22/07/22	55
57	Correcciones del asesor	15 días	lun 04/07/22	vie 22/07/22	55
58	Periodo de realizar las correcciones	5 días	lun 25/07/22	vie 29/07/22	57
59	Visto bueno del asesor	1 día	lun 01/08/22	lun 01/08/22	58
60	Entrega final de la tesis	1 día	mar 02/08/22	mar 02/08/22	59

Fuente: elaboración propia.

12. FACTIBILIDAD

Para la elaboración e implementación de este trabajo de graduación se cuenta con el permiso y acceso a la información en una aseguradora.

12.1. Factibilidad operativa

La factibilidad operativa se refiere al personal que se necesita para llevar a cabo el proyecto con éxito.

- Recurso humano necesario:
    - 1 persona con el conocimiento y experiencia necesaria para implementar la propuesta, con conocimientos de programación, OCR, Utilización de herramientas en la nube.
    - 1 persona para realizar el entrenamiento de la herramienta, las pruebas, la medición de confiabilidad y la medición del tiempo de los procesos
    - Se necesita de una persona para dar seguimiento al mantenimiento de la herramienta.

- Tiempo de observación

El tiempo de observación necesario será de 5 días, durante 2 horas al día para conocer el proceso sin la herramienta y paralelamente se implementará la herramienta y se entrenará hasta alcanzar el 95 % de confiabilidad.

- Tiempo de entrenamiento a la herramienta tecnológica

Después de crear la herramienta es necesario entrenarla para que aprenda a reconocer los datos y el tipo de documento que procesará, para iniciar se dará un entrenamiento de 50 tipos de documento y posteriormente se realizarán pruebas.

- Tiempo de pruebas tecnológicas

Para las realizar las pruebas tecnológicas será necesario que una persona observe la ejecución de la herramienta para verificar que la misma procesa los datos de manera correcta a través de realizar la medición de confiabilidad, si la confiabilidad es menor al 95% se debe volver a la etapa de entrenamiento como se mostró en el diagrama de entrenamiento y pruebas.

- Acceso a la información

Para la elaboración del presente trabajo es necesario acceder a la información relacionada con el reclamo de seguros la cual se encuentra en la base de datos de la empresa y los documentos escaneados se encuentran en una carpeta indexada utilizada por los sistemas de la aseguradora.

- Permisos

Permiso de la gerencia para realizar el proyecto lo cual brinda acceso al departamento de reclamos para la observación del proceso actual y el apoyo del departamento de informática para brindar la infraestructura y accesos necesarios para la creación de la herramienta.

La aseguradora dio su consentimiento para la elaboración del proyecto por lo cual se puede concluir que se realizará la implementación de dicho proyecto.

12.2. Factibilidad técnica

Evalúa si la empresa cuenta con la infraestructura técnica adecuada para crear e implantar la propuesta con éxito.

- Aspectos de Hardware
    - 1 computadora
    - Acceso a internet
- Aspectos de Software
    - Infraestructura de desarrollo en la empresa
- Conocimientos de pruebas técnicas y usuario final
    - Conocimientos en programación
- Tiempo necesario para el desarrollo de la solución
    - La propuesta se planificó para iniciar en enero del año 2022 y con una duración de 7 meses desde el inicio de la creación de la herramienta.

La empresa cuenta con experiencia en innovación de proyectos y cuenta con personal con la capacitación necesaria, por lo que se puede concluir que la

empresa cuenta con los recursos técnicos necesarios para realizar la propuesta con éxito.

12.3. Factibilidad económica

A continuación, se realizará una comparación de los costos que actualmente realiza la empresa y los costos que conlleva la herramienta y posteriormente se hará una comparación de ambos costos.

Descripción de costos para la creación y mantenimiento de la herramienta

- Computadora: con procesador core i5 de sexta generación, memoria ram al menos 8 GB, disco duro al menos de 1 T y teclado, mouse y monitor de 21” con un costo aproximado de Q8,000.00 este costo se presenta como inversión inicial.
- Internet: Costo de Q250.00 mensual, se va a estimar este costo mensual debido a que el servicio es un costo fijo usado en varios departamentos de la empresa por la misma razón no se colocará en la inversión.
- Programador: costos por hora Q100.00. Para la creación de la herramienta se estiman 2 horas diarias por 6 meses y luego para el mantenimiento 2 horas por semana.
- Servicios OCR: para obtener este costo estimado se supondrá una demanda de 3,000 reclamos al mes, se estima un costo por reclamo de $0.12 (Se aproxima a Q 1.00)
- Servicios Machine Learning: costo estimado por hora $1.00 (se aproximan a Q 8.00) y se estiman 25 horas de entrenamiento.

Tabla XIII. **Costos de creación de la herramienta**

No.	Costos	Cantidad	Costo por documento	Costo por hora	Mes	Año
	Costos iniciales					
1	computadora	----------	----------	----------	Q 666.67	Q 8,000.00
3	Programador para crear	240 horas	----------	Q 100.00	Q 4,000.00	Q 24,000.00
	Inversión inicial					**Q 32,000.00**
	Costos fijos					
2	Internet	----------	----------	----------	Q 250.00	Q 3,000.00
4	Programador para mantenimiento	2 horas /semana		Q 100.00	Q 800.00	Q 9,600.00
5	Servicios OCR	----------	Q 0.12	----------	Q 351.00	Q 4,212.00
6	Servicios Machine Lerning	----------	----------	Q 8.00	Q 200.00	Q 2,400.00
7	2 empleados				Q 25,992.48	Q 135,909.86
	Total					**Q 155,122.00**

Fuente: elaboración propia.

Descripción de costos como se realiza el proceso actualmente sin el uso de la herramienta.

En la tabla se colocó los costos que representa tener a 11 empleados en la planilla, actualmente se cuenta con este personal para atender los reclamos de seguro. Se supondrá que las personas ganan el sueldo mínimo que maneja la empresa que es de Q4,000.00.

Tabla XIV. **Costos de la planilla actual**

Numeral	Descripción	Mensual	Anual
1	Sueldo	Q 4,000.00	Q 48,000.00
2	Bono 14	Q 4,000.00	Q 4,000.00
3	Aguinaldo	Q 4,000.00	Q 4,000.00
4	IGSS cuota patronal del 10.67%	Q 426.80	Q 5,121.60
5	Vacaciones	Q 166.67	Q 2,000.00
6	Indemnización	Q 402.77	Q 4,833.33
	Costo total de 1 empleado	Q 12,996.24	Q 67,954.93
	Costo de 11 empleados	Q 142,958.64	Q 747,504.23

Fuente: elaboración propia.

El beneficio económico se obtendrá de la disminución de personal al utilizar la herramienta creada. Al poner en marcha la herramienta se eliminará por completo el personal asignado a las tareas de reclamos de seguros, quedando únicamente la persona que actualmente se encuentra a cargo. A continuación, se presenta una tabla con una comparación de costos:

Tabla XV. **Comparación de costos**

Año	Costos del proyecto	Costo situación actual costo anual	Diferencia anual
Inversión	Q 32,000.00		
1	Q 155,121.86	Q 747,504.23	Q 560,382.37
2	Q 155,121.86	Q 747,504.23	Q 592,382.37
3	Q 155,121.86	Q 747,504.23	Q 592,382.37
4	Q 155,121.86	Q 747,504.23	Q 592,382.37
5	Q 155,121.86	Q 747,504.23	Q 592,382.37
Total	**Q 807,609.30**	**Q 3,737,521.15**	**Q 2,929,911.85**

Fuente: elaboración propia.

Según la comparación de costos de ambos escenarios se obtiene un ahorro de Q 592,382.37 anual utilizando la herramienta tecnológica propuesta, por lo que se concluye que el proyecto cuenta con factibilidad económica.

REFERENCIAS

1. Ahirrao, S.; Baviskar, D.; Kotecha K. y Potdar V. (marzo, 2021). Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions. *IEEE Access 9,* 72894-72936. Recuperado de https://ieeexplore.ieee.org/document/9402739

2. Anchal, G. y Rishabh, M. (julio, 2020). Text extraction using OCR: A Systematic Review*. Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, Conferencia llevada a cabo en Coimbatore, India.

3. Ankit, B. (2017), *Document classification using machine learning* (Tesis de maestría). San José State University, San José California, Estados Unidos.

4. Eckert, K.; Montenegro, S.; López, N. y Candia, G. (septiembre, 2019). Extracción de Información de Evoluciones Clínicas Digitales mediante técnicas de Machine Learning. *X Congreso Argentino de Informática y Salud (CAIS).* Conferencia llevada a cabo en Sociedad Argentina de Informática e Investigación Operativa, Universidad Nacional de La Plata. Argentina.

5. Fonseca, J. (2019). *Improving OCR Post Processing with Machine Learning Tools* (Tesis de doctorado). University of Nevada, Las Vegas.

6. Jamshed, M.; Sami, M.; Rizwan, A. y Mueen, U. (julio, 2020). Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). *IEEE Access 8,* 142642-142668. Recuperado de https://ieeexplore.ieee.org/document/9151144

7. León, T. (2021). *Desarrollo de un sistema de inteligencia artificial para el reconocimiento de calificaciones manuscritas en exámenes* (Tesis de licenciatura). Universidad Politécnica de Cartagena, España.

8. Méndez, C.; Ordoñez, A.; Ordoñez, C.; Ordoñez, H. y Edier, A. (2019). *Sistema de Indexación de Documentos Jurisprudenciales Soportado en Inteligencia Artificial.* (Tesis de licenciatura). Fundación universitaria de Popayán, Popayán, Colombia.

9. Sainz, L. (2019). *Análisis de la aplicación de la inteligencia artificial para la mejora del proceso de gestión de facturas en una empresa industrial* (Tesis de licenciatura). Universidad de Cantabria, España.

10. Thomas, H. (23 de junio 2021). OCR with Tesseract, Amazon Textract, and Google Document AI: A Benchmarking Experiment. *Journal of Computational Social Science* 1, 1-38. Recuperado de https://link.springer.com/article/10.1007/s42001-021-00149-1