UNIVERSIDAD CÉSAR VALLEJO

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Sistema RPA utilizando técnicas de web scraping para garantizar la calidad de datos

TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

AUTORES:

Herrera Portella Luis Anthony (ORCID: 0000-0003-4032-9700)

Revilla Vergaray Franklin (ORCID: 0000-0002-0646-8507)

ASESOR:

Dr. Iván Carlo Petrlik Azabache (ORCID: 0000-0002-1201-2143)

LÍNEA DE INVESTIGACIÓN:

Sistemas de Información y Comunicaciones

LIMA – PERÚ

2021

Dedicatoria

Este proyecto se lo dedicamos a Dios, a nuestros familiares y personas especiales por darnos la fortaleza para cumplir este gran objetivo. A todas las personas especiales por su apoyo y compañía en esta trayectoria.

Agradecimientos

Nuestro profundo agradecimiento a Dios y familiares por haber sido mi apoyo durante todo este tiempo y a nuestro asesor por su apoyo durante todo el transcurso del taller.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El objetivo de esta investigación fue determinar la influencia de un sistema RPA utilizando técnicas de web scraping para garantizar la calidad de datos en la empresa Konecta, Lima 2021. La investigación tuvo un diseño preexperimental. Como instrumento, se utilizó la ficha de registro antes y después de la implementación del sistema RPA, guardando así el registro de los procesos involucrados para la extracción del conjunto de datos de cada día. Estas fichas de registro se encargaron de medir 2 dimensiones de la calidad de datos que consideramos adecuados para esta investigación: Consistencia con el indicador de Consistencia de formato de datos y Actualidad con su indicador de Frecuencia de actualización. Los resultados evidenciaron que la consistencia de formato de datos incrementó su porcentaje, teniendo como media del pretest = 80.65 y el postest = 100. Del mismo modo, se incrementó la media para frecuencia de actualización (pretest = 73.19; postest = 99.23). Se concluyó entonces que el RPA benefició enormemente al área de TI de la empresa, debido a que el personal que realizaba dicho proceso dispone de más tiempo para realizar otras actividades; lo cual validó nuestra hipótesis alterna y rechazó la hipótesis nula.

Palabras clave: Calidad de datos, web scraping, RPA, base de datos, SCRUM.

ABSTRACT

The objective of this research was to determine the influence of an RPA system using web scraping techniques to ensure data quality in the company Konecta, Lima 2021. The research had a pre-experimental design. As an instrument, the record card was used before and after the implementation of the RPA system, thus keeping the record of the processes involved for the extraction of each day's data set. These record cards were responsible for measuring 2 dimensions of data quality that we considered appropriate for this research: Consistency with the indicator of Consistency of data format and Actuality with its indicator of Update Frequency. The results showed that data format consistency increased its percentage, with the pretest mean = 80.65 and the posttest = 100. Similarly, the mean for update frequency increased (pretest = 73.19; posttest = 99.23). It was concluded then that the RPA greatly benefited the IT area of the company, due to the fact that the personnel performing this process have more time to perform other activities; which validated our alternative hypothesis and rejected the null hypothesis.

Keywords: Data quality, Web scraping, RPA, Database, SCRUM

I. INTRODUCCIÓN

El mundo actual presenta constantes cambios tecnológicos y las empresas están cada vez más interesadas en los datos como una herramienta esencial para la toma de decisiones correctas y oportunas, surgiendo así la obligación de que estos tengan la calidad necesaria. Un artículo de Veiga, Saraiva, otros (2017) menciona que los datos son de calidad si representan correctamente el mundo real y si satisfacen las necesidades del consumidor (pp.4-7). También surge la necesidad de optimizar tiempos y procesos de extracción de datos. Este contexto estará relacionado a garantizar la calidad de datos recolectados desde la web y será realizado por un RPA (Automatización robótica de procesos) utilizando técnicas de web scraping y así optimizar los procesos en las organizaciones.

En el ámbito internacional, Li Cai & Zhu (2015) en su artículo de la Data Science Journal, menciona que el big data presenta características de 4V (volumen, velocidad, variedad y valor), y estos son procesados y utilizados por las empresas; la extracción de datos se convierte en un problema urgente al no ser manejado de manera correcta (p.3). Por eso para una empresa es de suma importancia tener datos disponibles en tiempo y forma, porque con ello se lograría entender los últimos movimientos en el negocio y actuar a tiempo.

En Latinoamérica, una publicación en el repositorio de la Universidad de Antioquia - Colombia, indica que muchas compañías realizan de forma manual actividades que suelen ser muy repetitivas, invirtiendo así el tiempo y energía que deberían ser empleado para actividades de más valor (Serna ,2021, p.3). Para garantizar que los datos sean de calidad es necesario disminuir las tareas manuales repetitivas.

En el ámbito nacional, una encuesta de Transformación Digital (PAD-RTM, 2020, p.31), infiere que sólo el 13% de 324 empresas utilizan RPA. Por otro lado, según INEI en Resultados de la Encuesta Nacional de Empresas (2015, p.12) manifiesta que el 69,7% de las empresas consideran que la calidad de un producto o servicio es una de las estrategias más destacadas para posicionarse bien en el mercado.

Esta investigación se desarrolló en una empresa que presta servicio de call center. En la empresa, el área de TI existe personal que tiene la función de extraer de manera manual un conjunto de datos por día (el conjunto consta de 8 bases de datos que pasarán a llamarse ítems de datos) de la web corporativa. Estos datos son de la atención de los

clientes que se comunican al call center. Hay que mencionar que, como compromiso, cada ítem de datos de este conjunto debe ser guardado en el repositorio que le corresponde con el nombre y formato adecuado, estos deben estar disponibles antes de las 11 am.

El proceso de extracción de datos al ser una tarea manual y con frecuencia diaria genera las siguientes averías.

- Dejar las funciones principales para realizar la extracción de datos.
- En caso los datos no estén actualizados, se revisará 1 hora después
- Recolectar datos incompletos y/o erróneos.
- En oportunidades el encargado se olvida del proceso.
- Extraer los datos fuera del horario acordado.

Los puntos anteriores desprenden las siguientes consecuencias.

- Mala calidad del conjunto datos, con formato incorrecto o desactualizados.
- Las áreas interesadas no podrán realizar la toma de decisiones.
- El área de tecnología recibirá reclamos mediante tickets.
- Malestar y pérdida de credibilidad en el área de TI.

El siguiente gráfico muestra el porcentaje de cumplimiento de actualización y formato de cada ítem de cada conjunto de datos en el mes de julio del 2021. Evidenciando la necesidad de desarrollar un RPA para que el cumplimiento pueda ser al 100% cada día, ya que se trata de una fuente básico para la toma de decisiones.

Figura 1: Cumplimiento de actualización y formato de datos.

Id_conjunto	FECHA	Q ítems	Nro Item Cumple Formato	Nro Item Cumple Hora	% Cumple Formato	% Cumple Hora
20210701	1/07/2021	8	7	6	87.50%	75.00%
20210702	2/07/2021	8	8	8	100.00%	100.00%
20210703	3/07/2021	8	7	7	87.50%	87.50%
20210704	4/07/2021	8	6	5	75.00%	62.50%
20210705	5/07/2021	8	8	8	100.00%	100.00%
20210706	6/07/2021	8	6	5	75.00%	62.50%
20210707	7/07/2021	8	5	5	62.50%	62.50%
20210708	8/07/2021	8	6	6	75.00%	75.00%
20210709	9/07/2021	8	6	6	75.00%	75.00%
20210710	10/07/2021	8	6	6	75.00%	75.00%

Fuente: Elaboración propia de los autores.

Según la Figura 1, la Frecuencia de Actualización se ve afectado al no cumplir con la actualización de los datos con la periodicidad establecida; de la misma forma el indicador Consistencia de formato de datos, lo cual dificulta la lectura de estos al descargarse en un formato no permitido.

Si el problema persistía, entonces las áreas internas de la empresa no podrían realizar las actividades como retroalimentación, planificación, acciones correctivas. Esto conlleva a la pérdida de un cliente externo debido a no cumplir con las expectativas y al no lograr contrastar la información diaria, además de dejar en evidencia el incumplimiento y el poco uso de la tecnología para dar soluciones.

La propuesta de solución fue implementar un Sistema RPA que utilizando las técnicas de web scraping pudieran garantizar que los datos estén disponibles con calidad requerida según lo acordado. Este se encargaría de ingresar a la web, validar si los datos están actualizados y extraer el conjunto de datos que le corresponda a cada día, en el caso estos no estén actualizados; el RPA estará programado para ejecutarse en distintas horas antes de las 11:00 am, además una última ejecución a las 02:30 pm debido a que a veces los datos no están disponibles en el horario antes mencionado.

Seguidamente, el problema general se formula en forma de pregunta: "¿De qué manera influye un sistema RPA utilizando técnicas de web scraping en la garantía de la calidad de datos en la empresa Konecta, Lima 2021?", los específicos son: "¿En qué medida un sistema RPA utilizando técnicas de web scraping influye en la consistencia de formato de los datos para la garantía de la calidad de estos mismos, Lima 2021?", y "¿En qué medida un sistema RPA utilizando técnicas de web scraping influye en la frecuencia de actualización para la garantía de la calidad de datos en la empresa Konecta, Lima 2021?".

Adicionalmente, el proyecto se justificó en el aspecto teórico, debido a que se pretende adquirir nuevos conocimientos para este proyecto relacionados al proceso de extracción de datos con web scraping. De la misma manera se justificó en el aspecto social porque con esta investigación se le brindará apoyo a la empresa Konecta para mejorar su flujo de trabajo. Se justifica en el aspecto económico ya que este nuevo proceso al ser automático no se requerirá contratación de personal para esta actividad, por lo tanto, no incurrirá en gasto alguno. Además, el proyecto se justificó en el aspecto operativo dado que ahora este proceso de recolección será más eficiente y se logrará en el momento oportuno. Finalmente, se justificó en el aspecto tecnológico pues se implementará un sistema RPA para la optimización de la extracción de datos.

A continuación se precisa el objetivo general: determinar la influencia de un sistema RPA utilizando técnicas de web scraping para garantizar la calidad de datos en la empresa Konecta, Lima 2021, el primer objetivo específico es: determinar la influencia del sistema RPA utilizando técnicas de web scraping en la consistencia del formato de datos para garantizar la calidad de datos en la empresa Konecta, Lima 2021, y como segundo objetivo específico tenemos: determinar la influencia de un sistema RPA utilizando técnicas de web scraping en la frecuencia de actualización para garantizar la calidad de datos en la empresa Konecta, Lima 2021.

Finalmente, se plantea la siguiente Hipótesis principal: un sistema RPA utilizando técnicas de web scraping garantiza la calidad de datos en la empresa Konecta, Lima 2021, como hipótesis específica 1: un sistema RPA utilizando técnicas de web scraping aumenta el grado de consistencia de formato de los datos para la garantía de la calidad de datos en la empresa Konecta, Lima 2021, y como hipótesis específica 2: un sistema RPA utilizando técnicas de web scraping incrementa el grado de frecuencia de actualización para la garantía de la calidad de datos en la empresa Konecta, Lima 2021.

II. MARCO TEÓRICO

Con el transcurso del tiempo se han desarrollado diversos sistemas RPA y también usado técnicas de web scraping para la automatización de diferentes procesos; de la misma forma la calidad de los datos siempre ha sido un tema para tomar en cuenta en distintas empresas, es importante considerar algunos antecedentes y comprender las diferencias de estos.

Marshall Fernández (2018) en su tesis titulada: "Desarrollo de un modelo de calidad de datos aplicado a una solución de inteligencia de negocios en una institución educativa: CASO LAMBDA. Su problemática radica en que las soluciones de Inteligencia de Negocios son vulnerables a los problemas de consistencia que logren presentar los datos, debido a que los procedimientos realizados para la transformación conllevan a obtener reportes inconsistentes que generan problemas en la administración de la corporación. El objetivo general fue desarrollar un modelo de calidad de datos para una solución de Inteligencia de Negocios que se usa en dicha institución, esto está basado normas internacionales ISO/IEC. La investigación es de tipo aplicada y el diseño de tipo experimental puro. La población fue el Data Warehouse de la institución y su muestra fue LAMBDA.WAREHOUSE que son datos de algunas tablas elegidas, finalmente el resultado fue que los datos tienen un nivel de calidad bastante aceptable, particularmente en lo que concierne a la conformidad, completitud exactitud y credibilidad, aunque habiendo algunas observaciones en cuanto a su actualidad y consistencia".

Seguidamente Luis Felipe Jordan Conto Quispe y Nancy Margarita Rivera Quispe (2020) en su tesis titulada: "Aplicación móvil mediante RPA para la gestión de incidencias del área de soporte técnico. La problemática se centró en la demora en la atención y resolución de incidencias reportadas. El objetivo principal fue determinar el resultado que genera la aplicación móvil mediante RPA para gestionar incidencias en el área de soporte técnico. El tipo de investigación es de tipo aplicada y el diseño es pre-experimental. La población y muestra fueron de 60 incidencias que fueron resueltas durante 2 semanas, y así como resultado fue que el robot mejoró el porcentaje de incidencias solucionadas que se reportan a diario dicha área, habiendo ahora un promedio de 85.47% de incidencias resueltas; además se disminuyó significativamente el tiempo promedio de resolución de incidencias".

De la misma forma Rafael Velasco (2018) realizó la siguiente investigación titulada: "Sistema informático para el control de calidad de datos e información

estadística en los establecimientos de CLAS Batanes. Su problemática es que no existía algún filtro o control para la información que se registra en el sistema, obteniendo una serie de disconformidades con la información que se envía ya que esta no era exacta. El objetivo general es implementar un sistema informático para controlar la calidad de los datos de información estadística en las instalaciones. El tipo de investigación es aplicada y el diseño de tipo experimental puro". La población y muestra fue el personal administrativo del CLAS Batanes, el resultado fue que disminuyó a 21 fichas mal llenadas en promedio, se redujo a 21 minutos el tiempo empleado para control de calidad y se tiene un 60% de satisfacción por parte del usuario final con el sistema informático.

Existe también el trabajo de investigación de Brady Marlon Príncipe Quiñones y Christiam Alberth Mendoza Ruiz (2019) titulado: "Automatización robótica de procesos en las conciliaciones bancarias de una empresa industrial. La problemática fue que el área de Contabilidad no consigue entregar a tiempo los documentos contables de la información financiera en el cierre de cada mes, de tal forma afectando la operatividad de los demás procesos que exigen el traspaso oportuno de esta información para conciliar las cuentas bancarias. El objetivo principal fue determinar el impacto del proceso automatizado en las conciliaciones bancarias de dicha empresa. El tipo de investigación es aplicada y el diseño de tipo cuasi-experimental". La población fue compuesta por la cantidad de ejecuciones del proceso de conciliaciones bancarias y su muestra sólo por 16 ejecuciones, teniendo como resultado la reducción en un 11.59% el valor del costo, la disminución de un 27.80% del tiempo que se utilizaba para las conciliaciones bancarias y una mejora del 62.50% de la satisfacción según los resultados del robot.

Otro trabajo de investigación es el de Antonio Federico Martinez Nuñez (2020) titulado: "Automatización de web Scraping de los diarios de noticias para la empresa Isuri, San Martín de Porres. La problemática radica en la dificultad para poder obtener las últimas noticias web del día. El objetivo general fue desarrollar la automatización de web scraping de los diarios de noticias para la empresa Isuri, San Martin de Porres. El tipo de investigación es aplicada y su diseño de tipo experimental puro". La población y muestra fueron todos los colaboradores del área de sistemas de la empresa, donde el resultado fue la disminución de forma considerable en los recursos computacionales, presupuesto y errores del proceso de web scraping en la captura de noticias.

En el ámbito internacional tenemos a Diego Rodríguez García con su tesis titulada: “Desarrollo RPA para monitoreo de calidad de datos y generación de alertas. Su problemática se centra en que el área de Tecnología de Información (TI) de la empresa farmacéutica debe disminuir la inconsistencia que existe en la calidad de los datos, debido a que el proceso del negocio ha tenido errores en los datos, y en ocasiones requieren reprocesos dado el grado de dificultad. El objetivo general fue diseñar e implementar un mecanismo automatizado para realizar el monitoreo y generar alertas, teniendo como base la adaptación de un modelo metodológico, en características relacionadas con la calidad de los datos de las fuentes de información”. El tipo de investigación es aplicada y el diseño de tipo experimental puro. La población y muestra fueron 280 pedidos de venta, teniendo como resultado que este sistema RPA incrementó la propuesta de valor a la organización, siendo la Precisión la única dimensión con un margen de error por encima de los criterios de aceptación, deduciendo así que se tuvo un resultado positivo en las otras dimensiones de calidad de datos.

Tenemos además a Fernanda Soto Guerrero con su investigación titulada: “Análisis de la problemática asociada con la baja calidad de datos en los sistemas de información. La problemática radica en que los datos almacenados contienen errores, esto conlleva un esfuerzo para detectarlos y corregirlos, evidenciando así el tiempo y costo que representa. El objetivo general fue proponer un marco conceptual y atributos de calidad de datos. El tipo de investigación es aplicada y su diseño es experimental puro. La población fueron 53 miembros de las áreas de Gestión de Información estratégica, Dirección General académica y Operaciones el área de Dirección de Apropiación, la muestra estuvo compuesta por 31 miembros de las áreas de Gestión de Información estratégica, Dirección General académica y Operaciones, donde el resultado fue que las características más importantes a considerar y salvaguardar dentro es la Precisión y la Consistencia”, también cabe señalar que el factor que más influyente en la baja calidad de datos se muestra el origen de estos como el más mencionado, no obstante éste atributo se considera como uno de los menos significativos para garantizar la calidad de los datos.

De la misma forma, Hugo Alfredo Condori Yujra (2014) con su tesis titulada: “Web scraping para la obtención de información actualizada en internet con push notifications para smartphone. La problemática fue la necesidad de obtener información de forma automatizada, según los intereses que cada persona tiene, mejor dicho, realizar un seguimiento a las actualizaciones de la información. El objetivo principal fue

desarrollar una aplicación web/móvil para obtener la información actualizada de una página web, utilizando Web Scraping para la recolectar los datos y Push Notification para notificar al interesado". El tipo de investigación es aplicada y el diseño de tipo experimental puro. La población y muestra fueron sólo dos personas que utilizaron un smartphone con sistema operativo android con conexión a internet, teniendo como resultado una calificación del 91.45% de calidad total, lo que se entiende como la satisfacción que se lleva el usuario al interactuar con el software.

Por otro lado, Jorge Hernando Mendez Matamoros (2017), en su trabajo de investigación titulado: "Mejoramiento de calidad en conjuntos de datos abiertos basado en la aplicación de métricas de consistencia lógica. Su problemática radica en que los errores de calidad en los conjuntos de datos son bastante frecuentes. El objetivo principal fue diseñar una secuencia de reglas que admita crear métricas para una consistencia lógica y evaluar a cada conjunto de datos publicados en la plataforma distrital de datos abiertos y que facilite distinguir los registros que no están según las métricas de consistencia lógica y así conseguir el índice de calidad de acuerdo con cada métrica". El tipo de investigación es básica y el diseño de tipo experimental puro. La población y muestra está compuesta por conjuntos de datos de la plataforma distrital de datos abiertos, y como resultado se obtuvo que el índice de error del conjunto de datos está entre el 7% y el 19%. En la práctica esto significa que, de 40911 registros, entre 2864 y 7773 son erróneos.

Existe también el trabajo de investigación de Magaly Rincón Rodríguez (2019) quien realizó la siguiente investigación titulada: "Plan de gestión de calidad de datos para mejorar la oportunidad y pertinencia de la información de la oferta institucional en la dirección de apropiación del ministerio TIC. En su problemática se detectaron fallas en el rendimiento del proceso de los datos y esto conlleva a la falta de credibilidad y oportunidad en la información entregada. El objetivo principal fue establecer un plan para gestionar la de calidad de los datos adecuado a la Dirección de Apropiación, para desplegar control para la calidad, homologación y estandarización de los datos con el fin de optimar la pertinencia y oportunidad de la información". El tipo de investigación es básica y el diseño de tipo experimental puro. La población fue el área de Dirección de Apropiación y la muestra estuvo compuesta por 5 Bases de datos, 5 responsables directos de dichas bases de datos y 4 expertos que poseen relación con la oficina de TI respecto a la gestión de los datos, donde como resultado fue factible resaltar los beneficios de implementar el plan para la gestión de estos, y que durante un corto mediano y largo plazo

admita disminuir brechas para estos procesos del departamento de Apropiación, perfeccionando así los reportes de información y así tomar las decisiones correspondientes.

Definiendo conceptualmente los términos relacionados a las variables de estudio tenemos que un sistema RPA o en español Automatización Robótica de Procesos es una tecnología que pretende reducir la participación humana en las aplicaciones informáticas, fundamentalmente en las tareas cotidianas y repetitivas. (Doguc, 2020, p.476). Seguidamente Ravi Teja Yarlagadda en su publicación en The International Journal of Creative Research Thoughts (2018) indica que el objetivo más importante de la RPA es superar el desafío de la escalabilidad de la inteligencia humana, la duración que tarda un humano en completar una tarea es alto, y es por eso por lo que RPA busca cerrar esa brecha (p.366). Adicionalmente Wright y Bott en su artículo de la revista IDM (2018) mencionan que el principal motor para la implantación de muchos sistemas de RPA es el importante ahorro de costes que suponen, normalmente entre un tercio y una quinta parte del coste de un miembro del personal equivalente a tiempo completo (p.24). Con el aumento de la inteligencia artificial y la sofisticación de la automatización, las empresas encontrarán nuevas formas de cubrir las necesidades de los clientes, incluso antes de que el cliente las conozca (Wiley, 2018, p.57).

En cuanto a la técnica de web scraping existen varias definiciones como la de Lopez (2018) donde menciona que es el proceso de recoger datos de las páginas web mediante técnicas automatizadas (p.1). Actualmente, para acomodarse a una diversidad de situaciones, las tecnologías existentes de web scraping se han individualizado a partir de operaciones más pequeños, asistidos por humanos hasta la utilización de sistemas totalmente automatizados que son capaces de transformar sitios web completos en un conjunto de datos (Zhao, 2017, p.1). De igual importancia es definir que el objetivo principal e importante en este proceso es extraer información de sitios web diferentes y no estructurados y transformarla en estructurada (Saurkar, Pathare, Gode, 2018, p.363). El web scraping es cada vez más valioso como una forma de reunir y clasificar sin esfuerzo la gran cantidad de datos accesibles en la web, ya que estos están implantados dentro del diseño y estilo de los sitios y deben separarse minuciosamente para que sean valiosos (Nandwan, Mishra, Patil, Siddiqui, 2021, p.13).

Seguidamente hay que mencionar que la metodología utilizada para el desarrollo del sistema será SCRUM, ya que como menciona Schwaber y Sutherland (2020, p. 3) Scrum es un modelo ligero que ayuda a las personas, los equipos y las organizaciones a crear valor a través de soluciones adaptables a problemas complejos. (p.3). Otra definición, sería la de los hermanos Koi-Akrofi y Akwetey en su artículo de la International Journal of Software Engineering & Applications, donde se indica que Scrum es simplemente una técnica ágil de entregar productos iterativos e incrementales utilizando retroalimentación frecuente y toma de decisiones colaborativa (2019, p.29). También proporciona un proceso eficiente en casos de cambio de requisitos, porque los requisitos siempre cambian (Luiz, Da Rocha, Brendon, Da Silva, Barucke, 2019, p.107). Además, tener en cuenta que las estrategias que se requieren actualmente para lanzar productos están orientadas a la entrega rápida resultados tangibles, y de respuestas ágiles, flexibles y necesarias para operar en mercados de evolución constante. (Menzinsky, López, Palacio, 2016, p.12).

Tabla 1: Juicio de experto de la metodología de desarrollo.

Expertos	Scrum	XP	ICONIX
Mg. Danny Montoya	17	12	12
Dr. Francisco Manuel Hilario Falcon	18	17	17
Dr. Ivan Carlo Petrlik Azabache	18	13	16
	17.7	**14.0**	**15.0**

Fuente: Elaboración propia.

Considerando el resultado de la Tabla 1 que hemos obtenido del juicio de expertos, se tiene una calificación de 17.7 para SCRUM que es la metodología de desarrollo de software empelada en la presente investigación.

Con relación a la calidad de datos según la publicación de Vancauwenber, se puede considerar que los datos son de muy buena calidad si son idóneos para ser usados en un propósito o un contexto determinado, por ejemplo, en la toma de decisiones y/o

planificación (2019, p.1). De forma similar Azeroual (2017) menciona que la calidad de los datos suele definirse como la idoneidad de estos para su utilización en determinados objetivos de uso requeridos, que deben estar libres de errores, ser completos, correctos, actualizados y coherentes (p.83). No proporcionar datos de alta calidad a la organización ha traído varios problemas como malas decisiones debido a datos incorrectos, alto costo de funcionamiento e insatisfacción al cliente (Jaya, Ishak, Sidi, Affendey, A.Jabar, 2017, p.2647).

Además mencionar que la calidad de datos se divide en 15 características según la norma ISO/IEC 25012 (Ver Anexo 11) definiendo así un modelo general de calidad para aquellos que están representados de forma estructurada internamente de un sistema informático, estos son clasificados en dos grandes grupos; la calidad de datos inherente que es el grado con el que las características tienen el potencial de cubrir las necesidades fijas y necesarias cuando los datos son manejados bajo contextos definidas y la calidad de datos dependiente del sistema, que esto es el grado con el que la calidad de datos es alcanzada y preservada a través de un sistema informático cuando estos son utilizados bajo escenarios específicos (Universidad Nacional de Río Cuarto, 2020, p.696).

A continuación, presentamos dos dimensiones, con los indicadores que se alinean a las necesidades específicas de nuestra investigación. A través de la dimensión de Consistencia, con su indicador de consistencia de formato de datos y con la dimensión de Actualidad, con su indicador de frecuencia de actualización de datos, mediremos la descarga de las bases de datos en el formato requerido y a la hora determinada por la empresa.

Tabla 2: Métricas elegidas de la ISO/IEC 25024 para la evaluación de calidad (ISO/IEC 25012, 2015).

Dimensiones	Definición	Indicador	Fórmula
Consistencia	"Los datos están libres de contradicciones y son coherentes con el resto de los datos en un contexto específico de uso".	"Consistencia de formato de datos Consistencia del formato de datos de este ítem de datos".	X=A/B A = "Número de ítems de datos donde el formato es consistente según lo definido". B = "Número de ítems de datos para los cuales la consistencia de formatos debe estar definida".
Actualidad	"Los datos tienen un tiempo adecuado para el contexto específico en que se utilicen".	"Frecuencia de actualización Grado al cual los ítems de datos son actualizados con la frecuencia requerida".	X=A/B A =" Número de ítems de datos actualizados con la frecuencia requerida". B = "Número de ítems de datos que tienen un requerimiento de frecuencia de actualización".

Fuente: Elaboración propia.

III. METODOLOGÍA

3.1 Tipo y diseño de investigación

Esta investigación es de enfoque es cuantitativo, debido a que recopilaremos y analizaremos datos, de esta forma probaremos nuestras hipótesis (Hernández y Mendoza, 2018, p.5). La investigación es aplicada a problemas, situaciones y características propias, además no a desarrollar teorías y ser aplicada de manera inmediata.

La investigación es de tipo aplicada, dado que el proyecto busca desarrollar e implementar un sistema automatizado (RPA) que utilizando técnicas de web scraping se garantizará que los datos extraídos sean de calidad, con una estructura comprensible (formato csv o txt, etc). Para (Gabriel, 2017, p.155), la investigación aplicada, práctica o empírica, es caracterizada por que busca aplicar y utilizar los conocimientos que se adquirieron. Este tipo de investigación demanda un gran uso del conocimiento, descubriendo conocimientos y tecnologías nuevas.

El diseño de la investigación es prexperimental por que podemos identificar y medir las causas de un efecto, de acuerdo con (Hernández y Mendoza, 2018, p.141) para trabajar un grupo y darle tratamiento, y después aplicar una medición para analizar cuál es el nivel de grupo, no tiene una referencia previa del nivel del grupo.

3.2 Variables y operacionalización.

3.2.1 Variable

Variable Independiente: Sistema RPA utilizando técnicas de web scraping.

Variable Dependiente: Calidad de datos.

3.2.2 Matriz de Operacionalización de las variables

Tabla 3: Matriz de Operacionalización de las variables de la investigación.

Variable	Definición conceptual	Definición operacional	Dimensión	Indicador	Instrumento	Fórmula
Sistema RPA utilizando técnicas de web scraping	RPA es una tecnología que pretende reducir la intervención humana en las aplicaciones informáticas; web scraping es el proceso de recolectar datos contenidos en páginas web mediante técnicas automatizadas.	Un sistema RPA ayudará a garantizar la calidad de datos en la empresa Konecta, ya que al ser un proceso automatizado el rango de errores será mínima o nula y el tiempo de ejecución será mucho menor ya que las tareas repetitivas las ejecutará el mismo robot a través de la técnica de web scraping.				

Calidad de datos	La calidad de datos suele definirse como la idoneidad de los datos para su utilización en determinados objetivos de uso requeridos, que deben estar libres de errores, ser completos, correctos, actualizados y coherentes	A través de la ficha de registro podemos medir esta variable ya que nos brindará data actualizada a una hora determinada y en el formato requerido, para que así la empresa pueda usarla a fin de tomar mejores decisiones.	Consistencia	Consistencia de formato de datos	Ficha de registro	X=A/B
						A = # ítems de datos donde el formato es consistente según lo definido.
						B = # ítems de datos para los cuales la consistencia de formatos debe estar definida.

			Actualidad	Frecuencia de actualización	Ficha de registro	X=A/B
						A = # ítems de datos actualizados con la frecuencia requerida.
						B = # ítems de datos que tienen un requerimiento de frecuencia de actualización.

3.3 Población, muestra y muestreo

3.3.1 Población.

Conjunto de casos definidos, limitados y accesibles que cumple criterios determinados y formará el referente para elegir la muestra (Arias, Villasís y otros, 2016, p.2), en efecto, la población para este estudio será de 31 conjuntos de datos de atención al cliente.

3.3.1.1 Criterios de inclusión

Cada uno de los 31 conjuntos de datos corresponde a los 31 días calendarios y está compuesto por 8 ítems de datos de atención a clientes.

3.3.1.2 Criterios de exclusión

El RPA no extraerá datos más allá de lo preestablecido, ya que en la web existen más datos de atención a clientes.

3.3.2 Muestra

Es un subconjunto de la población, constituida por unidades específicas de análisis, en este escenario la parte estudiada o muestra representa al total de la población, por tanto, no es imprescindible aplicar la fórmula de muestra (Hernández y Mendoza, 2018, p.135).

3.3.3 Muestreo

De acuerdo con (Hernández y Mendoza 2018, p.175) quien menciona que las muestras probabilísticas son cifras o componente que explican las propiedades de una población, para este contexto se aplicará estudio probabilístico aleatorio simple, debido a que conocemos la población total, además esta es pequeña.

Tabla 4: Tamaño de la muestra de la investigación.

Muestra	Periodo	Estratificado	Indicador
31 conjuntos de datos	1 mes	31 conjuntos de datos	Consistencia de formato de datos
31 conjuntos de datos	1 mes	31 conjuntos de datos	Frecuencia de actualización

Fuente: Elaboración propia

3.4 Técnicas e instrumentos de recolección de datos, validez y confiabilidad

Como instrumento, se tendrá a la ficha de registro pretest y postest en la implementación del sistema RPA, este último guardará en una tabla de SQL Server el detalle (fecha, hora

y formato) de cada ítem de datos extraído de la web, a partir de ello podemos completar la ficha de registro planteada para esta investigación.

3.4.1 Ficha de registro.

Se realizó la ficha de registro en el área de TI de la Empresa, para lo cual se utilizaron 2 fichas de registro para evaluar cada indicador (Ver anexo 4,5,6 y 7).

- FR1: Ficha de registro para el pretest.
- FR2: Ficha de registro para el postest.

Tabla 5: Técnica e instrumentos de recolección de datos.

Indicador	Técnica	Instrumento	Fuente	Informante
Porcentaje de consistencia de formato de datos	Fichaje	Ficha de registro	Registros del área de TI de la Empresa.	Área de TI
Porcentaje de frecuencia de actualización				

Fuente: Elaboración propia.

3.5 Procedimientos.

Se usará las fichas de registros log para los indicadores "consistencia de formato de datos" y "frecuencia de actualización" que serán evaluadas en la empresa Konecta Perú.

Se utilizarán dos fichas de registros para ambos indicadores, una ficha pretest (ver Anexos 4 y 5) y una ficha post test (ver Anexos 6 y 7) y así realizar una comparativa luego de la implementación del sistema.

Los registros log post test serán recolectados por el sistema, donde se observará la fecha y hora de extracción de cada ítem del conjunto de datos y también que éstas tengan el formato requerido, esta información será almacenada en una tabla de SQL.

3.6 Método de análisis de datos.

La interpretación se realiza teniendo en consideración cada uno de los niveles de medida para las variables y a través de la estadística inferencial y descriptiva. Se efectúa tomando como base la matriz de datos construida en un programa computacional (Hernández y Mendoza, 2018, p.311). Para el desarrollo de esta investigación nos centraremos en los aspectos descriptivos e inferenciales a fin de comprobar las hipótesis específicas

propuestas anteriormente. El proceso de los datos tomados de la muestra se realizó a través del software estadístico SPSS 25.0.

Estadística descriptiva

Los cálculos realizados fueron trasladados a tablas que se complementaron con gráficos para un mayor entendimiento de los resultados. Dichos cálculos fueron los siguientes:

- Mínimo y Máximo
- Media aritmética
- Desviación estándar

Estadística Inferencial

Las pruebas validaron las hipótesis planteadas en las dos dimensiones correspondientes. Dichas pruebas realizadas fueron:

- Prueba de Normalidad Shapiro-wilk
- Prueba de Wilcoxon
- Coeficiente de correlación de Pearson

3.7 Aspectos éticos.

"La Ley N. o 28858 autoriza al Colegio de Ingenieros del Perú inspeccionar a los profesionales de Ingeniería de la República y cuidar para que estas actividades sean desarrolladas dentro de las normas de ética profesional. Respetar el derecho a la autoría de la producción y obras de sus colegas docentes y de sus alumnos, evitando dar uso en beneficio propio o de terceros las investigaciones, los estudios, las tesis y demás trabajos realizados". (Codigo de Ética CIP, pp.1-12)

En el artículo 6°, se expone la honestidad y transparencia de la investigación de la misma manera el respeto como también el citado de trabajos. "Los investigadores convendrán asegurar que la investigación se ha elaborado cumpliendo rigurosamente con los requisitos legales, éticos y de seguridad, respetando los términos y condiciones determinadas en los proyectos de investigación". (Resolución de Consejo Universitario, 2017).

IV. RESULTADO

En esta sección se expondrán y detallarán los resultados alcanzados en esta investigación, así también dejaremos en evidencia, de manera certera y confiable los puntos previamente detallados

4.1. Análisis Descriptivo

En esta investigación, se desarrolló e implementó un sistema RPA utilizando técnicas de web scraping y así garantizar la calidad de los datos, por tanto, se aplicó un pretest que nos permitió conocer la problemática que concierne a los indicadores. Además, se realizó un postest para conocer los efectos del sistema desarrollado sobre el problema presentado.

A continuación, los resultados descriptivos.

- INDICADOR: Consistencia de formato de datos.

A continuación, presentamos los resultados descriptivos lo cuales fueron obtenidos del indicador consistencia de formato de datos se basaron en los datos recolectados de la ficha de registros que se encuentran en los Anexos 4 y 6.

Tabla 6: Medidas descriptivas de la Consistencia de formato de datos antes y después de implementar el Sistema RPA.

ESTADÍSTICOS DESCRIPTIVOS					
	N	**Mínimo**	**Máximo**	**Media**	**Desviación**
CFD_pretest	31	.625	1.00	.8064	.1062
CFD_postest	31	1.00	1.00	1.00	.0000
N válido (por lista)	31				

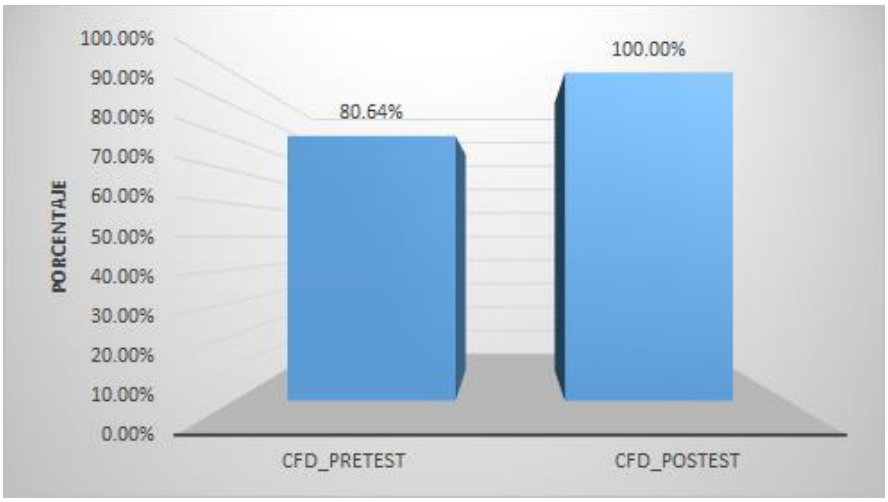
Fuente: Elaboración propia

En el indicador consistencia del formato de datos, dentro en el pretest obtuvo una media de 80,64% de conjunto de datos que en su extracción no cumplieron con el formato correspondiente. Asimismo, el resultado postest muestra que la media es de 100% evidenciando así que todos los conjuntos de datos son extraídos con el formato requerido. (ver Figura 2).

AL realizar la desviación estándar, para el pretest se presentó una variabilidad de 10.62%; sin embargo, en el postest se tuvo un valor de 0.00%.

Finalmente se obtuvo un mínimo de 62.50% en el pretest y un 100.00% en el postest, en cuanto al máximo se obtuvo un 100% tanto en el pretest como en el postest.

Figura 2: Porcentaje de la Media de Consistencia de formato de datos antes y después de implementar el Sistema RPA.



Fuente: Elaboración propia

- INDICADOR: Frecuencia de Actualización

En la Tabla 7 se pueden observar los resultados descriptivos para la frecuencia de actualización de los datos.

Tabla 7: Medidas descriptivas de la Frecuencia de Actualización antes y después de implementar el Sistema RPA.

ESTADÍSTICOS DESCRIPTIVOS					
	N	**Mínimo**	**Máximo**	**Media**	**Desviación**
FDA_pretest	31	.50	1.00	.7319	.1446
FDA_postest	31	.88	1.00	.9923	.0312
N válido (por lista)	31				

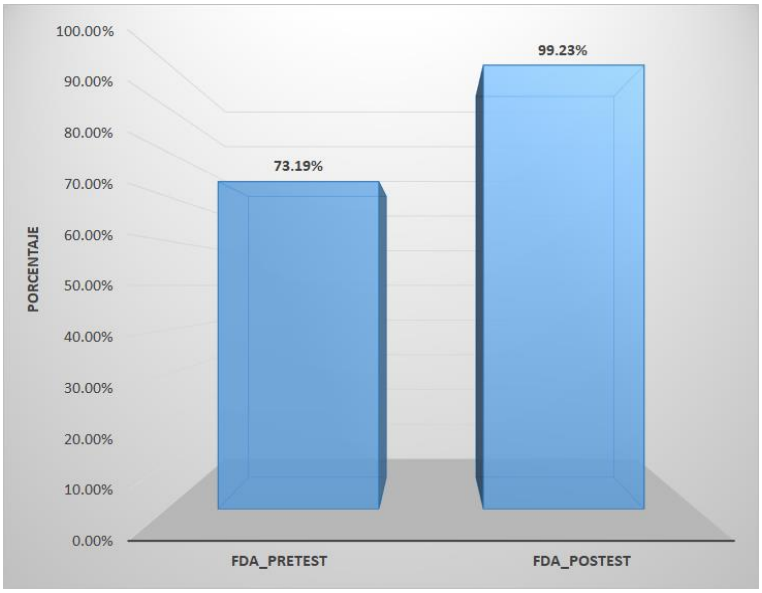
Fuente: Elaboración propia

Con relación al indicador Frecuencia de Actualización que está asociado a calidad de datos, para el pretest se obtuvo un valor de 73.19%, entre tanto que, para el postest fue del 99.23% así como se muestra en la figura 3; esto evidencia la gran variación antes y después de implementar del Sistema RPA.

Referente a la desviación estándar, para el pretest se presentó una variación de 14.46%; sin embargo, en el postest se tuvo un valor de 3.12%.

Finalmente se obtuvo un mínimo de 50% en el pretest y un 88% en el postest, en cuanto al máximo se obtuvo un 100% tanto en el pretest como en el postest.

Figura 3: Porcentaje de la Media de Frecuencia de Actualización antes y después de implementar el Sistema RPA.



Fuente: Elaboración propia

4.2. Prueba de normalidad

Se ejecutó la prueba de normalidad a cada uno de los indicadores, Consistencia del formato de datos y Frecuencia de Actualización a utilizando la prueba de Shapiro-Wilk, considerando que la muestra está compuesta por 31 conjunto de datos, siendo el valor inferior a 50, del mismo modo menciona Droppelmann (2018, p. 41). La prueba se efectuó cargando los datos para cada indicador en el software estadístico SPSS 25.0, para el nivel de confiabilidad del 95%, teniendo en cuenta las siguientes condiciones:

Si:

Sig. < 0.05 adopta una distribución no normal.

Sig. > 0.05 adopta una distribución normal.

Dónde:

Sig.: P-valor o nivel crítico del contraste.

A continuación, presentamos los resultados:

- INDICADOR: Consistencia del formato de datos

Con el fin de obtener la prueba de hipótesis hemos comparado la distribución de los datos, puntualmente si los datos del indicador Consistencia del formato de datos presentaban una distribución normal.

Tabla 8: Prueba de normalidad para la Consistencia de formato de datos antes y después de implementado el Sistema RPA.

	Shapiro-Wilk		
	Estadístico	**gl**	**Sig.**
CFD_pretest	**0.795**	**31**	**0.000**
CFD_postest		**31**	

Fuente: Elaboración propia

Según se aprecia en la Tabla 8, los resultados de manifiestan que el Sig. de Consistencia del formato de datos en el pretest fue de 0.00, dicho valor es menor que 0.05 por lo cual no se distribuye de forma normal. No se muestra los resultados de la prueba del Post-Test debido a que todos los resultados cuentan con un valor de 100% infiriendo que el Sig. de Consistencia de formato de datos que fue de 0.000, y de igual forma que el Pre-Test el valor es inferior a 0.05, por lo cual también se infiere que los datos tampoco se distribuyen de forma normal. Esto confirma que los datos analizados en ambos escenarios no tienen una distribución normal, esto se puede evidenciar en las Figuras 4 y 5.

Figura 4: Prueba de normalidad de la Consistencia de formato de datos antes de implementar el sistema RPA.

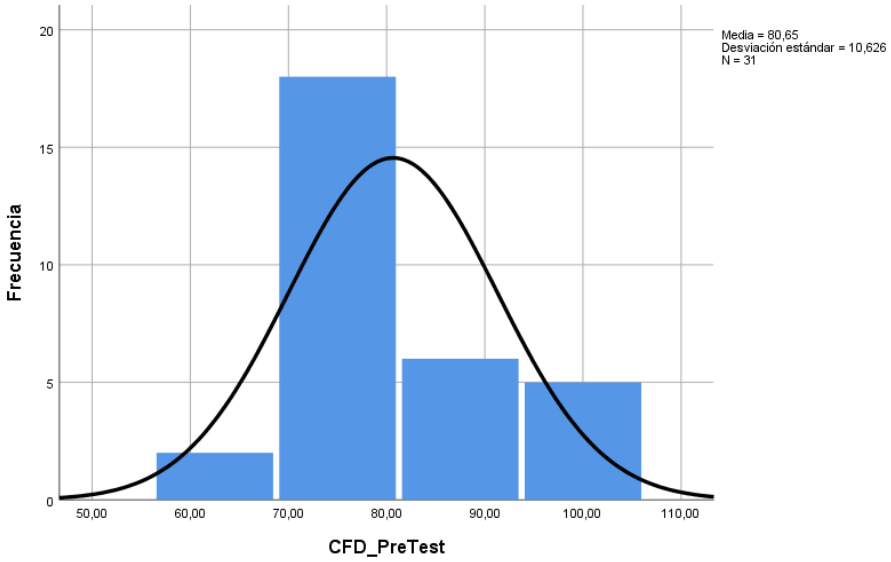
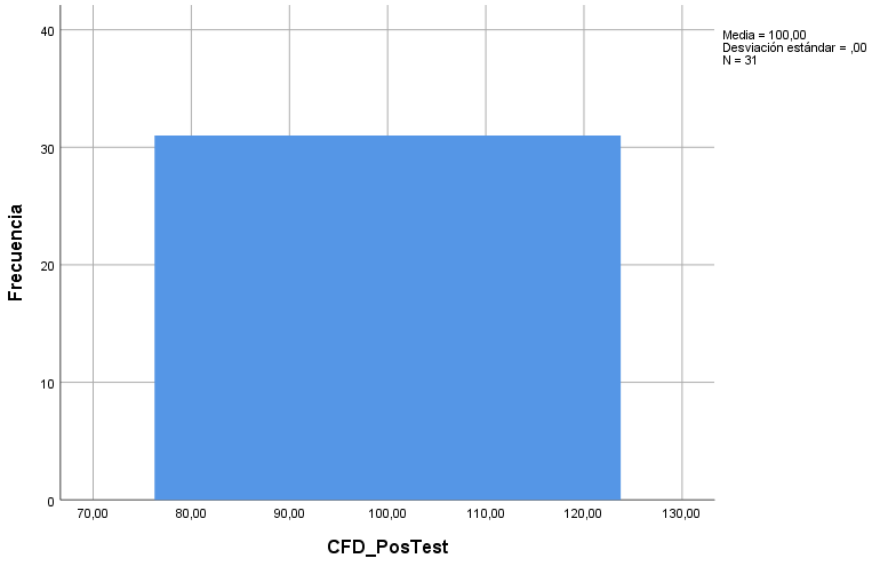


Figura 5 : Prueba de normalidad de la Consistencia de formato de datos después de implementar el sistema RPA.



- INDICADOR: Frecuencia de Actualización

Con la finalidad de obtener la prueba de hipótesis hemos comparado la distribución de los datos, puntualmente si los datos del indicador Consistencia del formato de datos presentaban una distribución normal.

Tabla 9: Prueba de normalidad para la Frecuencia de actualización antes y después de implementado el Sistema RPA.

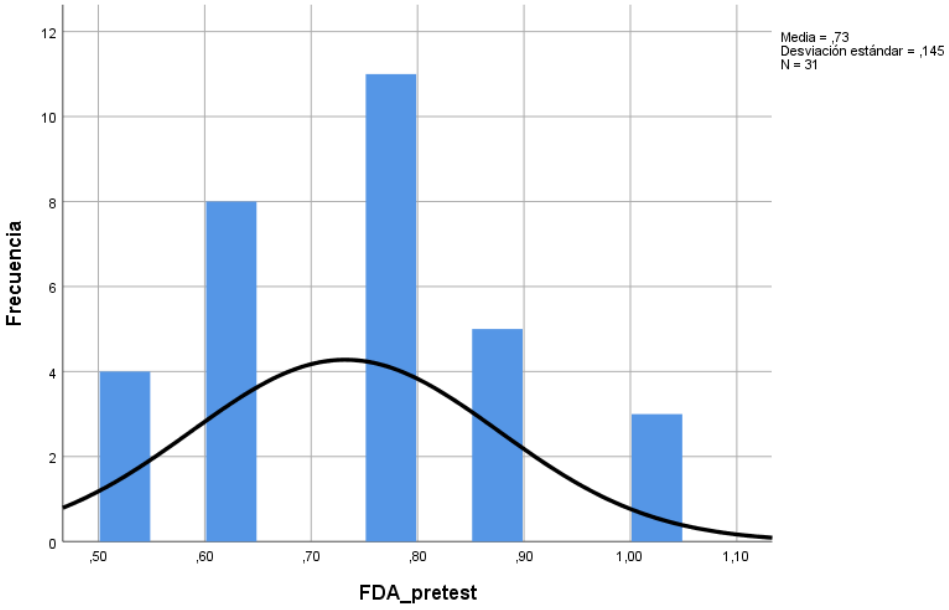
	Shapiro-Wilk		
	Estadístico	**gl**	**Sig.**
FDA_pretest	0.918	31	0.020
FDA_postest	0.270	31	0.000

Fuente: Elaboración propia

Se puede apreciar en la Tabla 9, los resultados obtenidos muestran que el Sig. de Frecuencia de actualización en la calidad de datos, tuvo en el pretest un 0.020, este valor es menos de 0.05, por esto llegamos a la conclusión que los datos no se distribuyen normalmente.

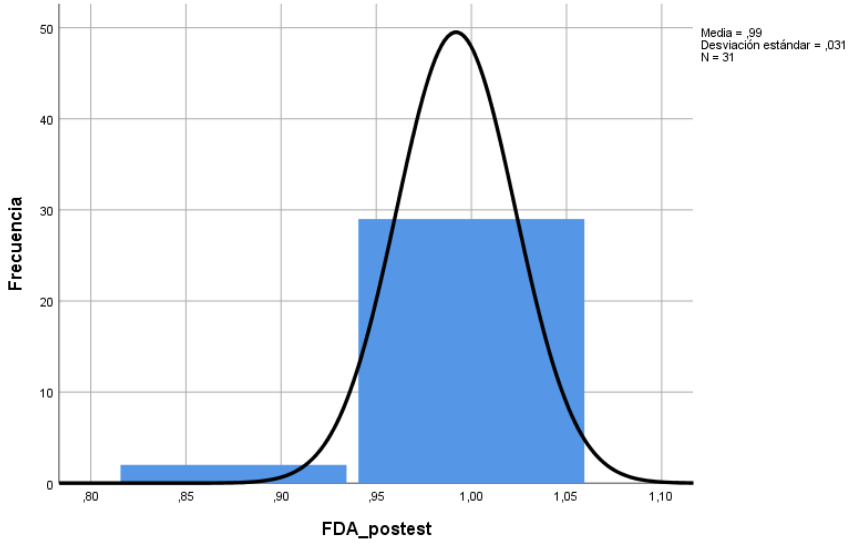
Los resultados de la prueba del postest indican que el Sig. de la Frecuencia de actualización fue de 0.000, y de igual forma que el pretest el valor es menor a 0.05, en efecto también se concluye que los datos tampoco se distribuyen de froma normal. Se puede confirmar que los datos analizados en ambos escenarios no siguen una distribución normal, lo cual se muestra en las Figuras 6 y 7.

Figura 6: Prueba de normalidad de la Frecuencia de actualización antes de implementar el sistema RPA.



Fuente: Elaboración propia

Figura 7: Prueba de normalidad de la Frecuencia de actualización después de implementar el sistema RPA.



Fuente: Elaboración propia

4.3. Prueba de Hipótesis

Hipótesis Específica 1

H1: Un sistema RPA utilizando técnicas de web scraping aumenta la consistencia de formato de datos y así garantizar la calidad de estos en la empresa Konecta.

Donde:

CFDa: Consistencia de formato de datos antes de implementar el RPA utilizando técnicas de web scraping.

CFDd: Consistencia de formato de datos después de implementar el RPA utilizando técnicas de web scraping.

Hipótesis H0: El sistema RPA utilizando técnicas de web scraping no aumenta la consistencia de formato de datos para la garantía de la calidad de datos en la empresa Konecta.

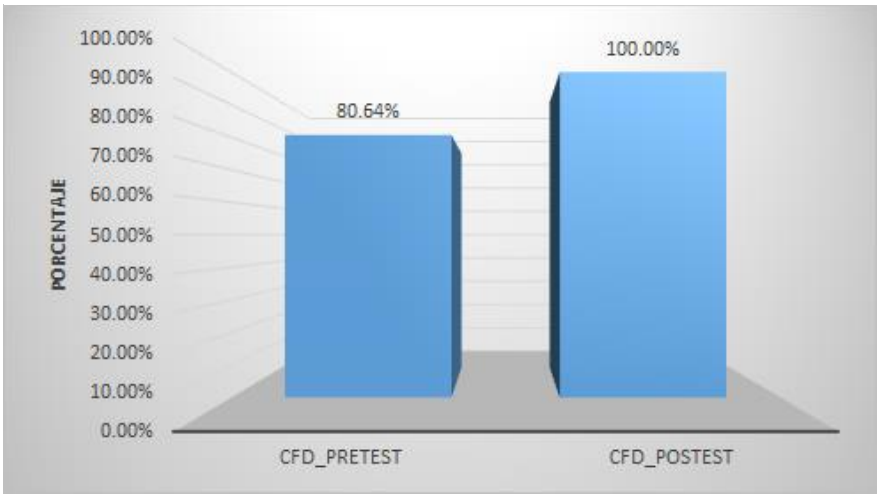
$$H0: RIDd \leq RIDa$$

Hipótesis Ha: El sistema RPA utilizando técnicas de web scraping aumenta la consistencia del formato de datos para la garantía de la calidad de datos en la empresa Konecta.

$$Ha: RIDd > RIDa$$

En la Figura 8, la consistencia de formato de datos en el Pre-Test es 84.64 % y en el Post-Test es 100%.

Figura 8 : Consistencia de formato de datos - Comparativa general de las medias.



Fuente: Elaboración propia

De acuerdo con la Figura 8, se presenta un crecimiento en el porcentaje de la consistencia del formato de datos, esto lo podemos verificar al contrastar las medias correspondientes, lo cual incrementa de un 80.64% al valor de 100%.

Con respecto al resultado de la prueba de hipótesis se utilizó la prueba no paramétrica de Wilcoxon, puesto que los datos conseguidos en la investigación no tiene distribución normal. El valor de Sig o P valor es 0.000, es menor que 0.05 (Ver tabla 10).

Tabla 10: Prueba de Wilcoxon para la Consistencia de formato de datos antes y después de implementar el Sistema RPA.

		RANGOS		
		N	**Rango promedio**	**Suma de rangos**
CFD_postest - CFD_pretest	Rangos negativos	0[a]	,00	,00
	Rangos positivos	26[b]	13,50	351,00
	Empates	5[c]		
	Total	31		

a. CFD_postest < CFD_pretest
b. CFD_postest > CFD_pretest
c. CFD_postest = CFD_pretest

ESTADÍSTICOS DE PRUEBA[a]	
	CFD_postest - CFD_pretest
Z	**-4,650[b]**
Sig. asintótica(bilateral)	**0.000**

a. Prueba de rangos con signo de Wilcoxon
b. Se basa en rangos positivos

Fuente: Elaboración propia

Seguidamente, se pasa a rechazar la hipótesis nula, por tanto, se acepta la hipótesis alterna teniendo un 95% de confianza. Esto quiere decir que, el sistema RPA utilizando técnicas de web scraping aumenta la consistencia del formato de datos para la garantía de la calidad de datos en la empresa Konecta.

Hipótesis Específica 2

H2: Un sistema RPA utilizando técnicas de web scraping incrementa el grado de frecuencia de actualización para la garantía de la calidad de datos en la empresa Konecta.

Donde:

NFAa: Grado de frecuencia de actualización antes de implementar el sistema RPA utilizando técnicas de web scraping.

NFAd: Grado de frecuencia de actualización después de implementar el sistema RPA utilizando técnicas de web scraping.

Hipótesis H0: El sistema RPA utilizando técnicas de web scraping no incrementa el grado de frecuencia de actualización para la garantía de la calidad de datos en la empresa Konecta.

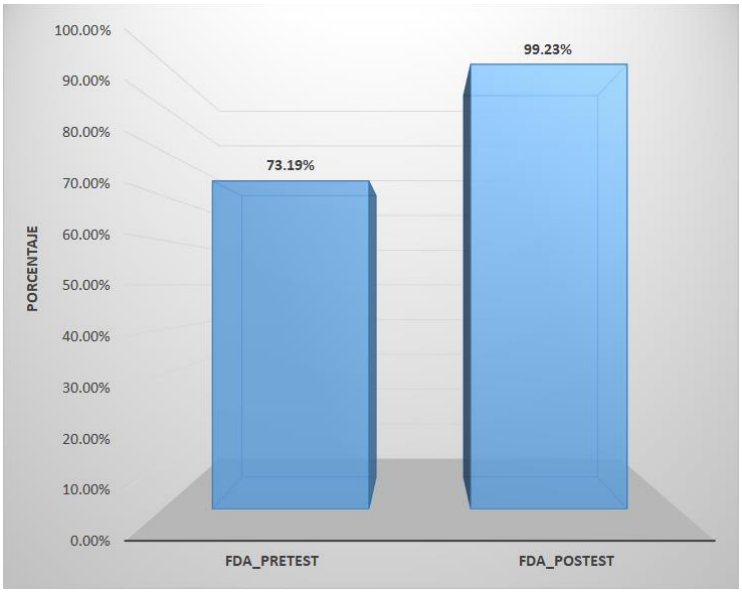
$$H0: NFAd \leq NFAa$$

Hipótesis Ha: El sistema RPA utilizando técnicas de web scraping incrementa el grado de frecuencia de actualización para la garantía de la calidad de datos en la empresa Konecta.

$$Ha: NFAd > NFAa$$

En la Figura 9, la Frecuencia de actualización en el pretest es de 73.19% y en el postest es de 99.23%.

Figura 9: Frecuencia de Actualización - Comparativa general de las medias.



Fuente: Elaboración propia

Para concluir, en la Figura 9 existe un crecimiento en el porcentaje de la Frecuencia de actualización de datos, esto lo podemos verificar si comparamos los valores respectivos, que van de un 73.19% alcanzando el valor actual de 99.23%.

Respecto al resultado de la prueba de hipótesis se aplicó la prueba no paramétrica Wilcoxon, ya que los datos conseguidos durante la investigación (Pretest y Postest) no pertenecen a la distribución normal. El valor de Sig o P valor es 0.000, el cual es visiblemente menos de 0.05 (Ver tabla 11).

Tabla 11: Prueba de Wilcoxon para la Frecuencia de Actualización en el proceso de calidad de datos antes y después de implementar el Sistema RPA.

		RANGOS		
		N	**Rango promedio**	**Suma de rangos**
FDA_postest - FDA_pretest	Rangos negativos	1[a]	1,00	1,00
	Rangos positivos	27[b]	15,00	405,00
	Empates	3[c]		
	Total	31		

a. FDA_postest < FDA_pretest
b. FDA_postest > FDA_pretest
c. FDA_postest = FDA_pretest

ESTADÍSTICOS DE PRUEBA[a]	
	FDA_postest - FDA_pretest
Z	**-4,648[b]**
Sig. asintótica(bilateral)	**.000**

a. Prueba de rangos con signo de Wilcoxon
b. Se basa en rangos negativos

Fuente: Elaboración propia

Por tanto, se rechaza la hipótesis nula, y se procede aceptar la hipótesis alterna teniendo un 95% de confianza. Por ese motivo, el sistema RPA utilizando técnicas de web scraping incrementa el grado de frecuencia de actualización para la garantía de la calidad de datos en la empresa Konecta.

IV. DISCUSIÓN

En este capítulo, se compararon los resultados de los antecedentes del marco teórico y los resultados obtenidos en esta investigación. Se mostrará las diferencias y semejanzas de las mediciones aplicadas a los indicadores, determinando así que la implementación del Sistema RPA utilizando técnicas de web scraping garantiza la calidad de datos.

En esta investigación, uno de los resultados conseguidos es antes de la implementación del Sistema RPA se tenía alrededor de 80.64% de porcentaje en la consistencia de datos (Consistencia), mientras que luego de realizar la implementación se tiene un 100%. Del mismo modo para la frecuencia de actualización (Actualidad), se incrementó la media de un 73.19% a un 99.23% luego de la implementación del Sistema RPA. Por esta razón, podemos afirmar que luego de implementar el Sistema RPA utilizando técnicas de web scraping para garantizar la calidad de datos en la empresa Konecta presenta un resultado positivo en el porcentaje de las dimensiones de Consistencia y Actualidad. De igual manera este resultado es similar al trabajo de Rodriguez (2020) en la cual su investigación "Desarrollo RPA para monitoreo de calidad de datos y generación de alertas" también plantea el efecto algunas dimensiones: Precisión, Completitud, que están relacionadas con el tiempo (Actualidad) y Consistencia. El RPA realizó proceso y generó alertas definidas en el universo cargado con la información del negocio, y se alertó aquellas dimensiones que poseían errores. Durante este proceso, se determinó que la única dimensión con errores superiores a los criterios de aceptación es la precisión, de esta manera concluyendo que se tuvo un resultado positivo en la dimensión de Actualidad (Relacionada con el tiempo) y Consistencia.

Caso contrario ocurre en la investigación de Fernandez (2018) donde podemos exponer en términos generales que los datos poseen una altura en cuanto a calidad bastante aceptable, no obstante, hay algunas fallas en su consistencia semántica y por la cantidad de llaves que tiene hay la posibilidad de que un riesgo altísimo de inconsistencia de datos generado por duplicidad, un contexto que se puede presentar al tener un mal manejo los datos.

Además, la característica más baja de la calidad de datos está en la actualización de esta, ya que es afectada de manera directa por las caídas de los procesos ETL, teniendo como resultado que los datos no estén actualizado en tiempo y forma según la necesidad del usuario, pese a que existen obligaciones de su actualización.

v. CONCLUSIONES

Después de implementar el sistema RPA utilizando técnicas de web scraping se puede afirmar que:

El RPA mejoró la calidad del conjunto de datos que son extraídos a diario, además se puede afirmar que se logró alcanzar un 100% en el indicador consistencia del formato de datos, logrando una fácil lectura de estos.

El uso del RPA mejoró drásticamente la frecuencia de actualización llegando al 99.19% y cumpliendo así con las expectativas del cliente y consumidores de dichos datos.

Las métricas fueron tomadas como base principal de la ISO/IEC 25024 sobre el espacio definido de la organización.

Podemos concluir que el RPA beneficia enormemente al área de TI de la empresa, debido a que el personal que realizaba dicho proceso tendrá más tiempo para realizar otras actividades, como análisis y mejora de otros procesos.

VI. RECOMENDACIONES

Se recomienda analizar a detalle las necesidades y los puntos resaltantes para medir la calidad en los datos y así evitar discordancias con las personas interesadas en estos, por posibles cambios no contemplados.

Derivar medidas de las dimensiones de la calidad de datos especificadas en la ISO/IEC 25012.

El sistema RPA utilizando técnicas de web scraping fue desarrollado para un entorno en específico y desde el código fuente está desarrollado y configurado para trabajar únicamente en éste, debido a características especiales como el acceso a una red privada.

Las empresas quieran desarrollar e implementar un RPA, deben disponer de manera constante de personal capacitado para dar mantenimiento o mejoras debido al constante avance de la tecnología como también según las necesidades.

REFERENCIAS

ARIAS-GÓMEZ, Jesús; VILLASÍS-KEEVER, Miguel Ángel; NOVALES, María Guadalupe Miranda. El protocolo de investigación III: la población de estudio. Revista Alergia México, 2016, vol. 63, no 2, p. 201-206, recuperado de https://www.redalyc.org/pdf/4867/486755023011.pdf

AZEROUAL, Otmane. Improving the Data Quality in the Research Information Systems. International Journal of Computer Science and Information Security (IJCSIS), 2017, vol 15, no 11. DOI: https://dspacecris.eurocris.org/bitstream/11366/633/1/Azeroual_IJCSIS_201711.pdf

CAI, L and ZHU, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 2015, vol 14, no 2. DOI: http://dx.doi.org/10.5334/dsj-2015-002

CIP, CÓDIGO DE ÉTICA DEL COLEGIO DE INGENIEROS DEL PERÚ, 2018, recuperado de http://www.cip.org.pe/publicaciones/reglamentosCNCD2018/codigo_de_etica_del_cip.pdf

CONDORI YUJRA, Hugo Alfredo. Web Scraping para la obtención de información actualizada de Internet con push notifications para smartphone. Tesis Doctoral. 2014, Recuperado de https://repositorio.umsa.bo/handle/123456789/8405

CONTO QUISPE, Luis Felipe Jordan; RIVERA QUISPE, Nancy Margarita. Aplicación móvil mediante RPA para la gestión de incidencias del área de soporte técnico. 2020, recuperado de https://repositorio.ucv.edu.pe/handle/20.500.12692/63937

DOGUC, Ozge. Robot Process Automation (RPA) and Its Future. Advances in E-Business Research, 2020, p.469-492. DOI: https://www.researchgate.net/publication/338302068_Robot_Process_Automation_RPA_and_Its_Future

DROPPELMANN, G. Pruebas de normalidad. Revista actualizaciones clínicas meds, 2018, vol. 2, no 1, p. 39-43, recuperado de https://meds.b-cdn.net/wp-content/uploads/Revista_cientifica_3.pdf

FERNÁNDEZ SÁENZ, Marshall André. Desarrollo de un modelo de calidad de datos aplicado a una solución de inteligencia de negocios en una institución educativa: Caso Lambda. 2018, Recuperado de https://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/12014

GABRIEL-ORTEGA, Julio. Cómo se genera una investigación científica que luego sea motivo de publicación. J. Selva Andina Res. Soc. [online]. 2017, vol.8, n.2 [citado 2021-07-31], pp. 155-156. Disponible en: http://www.scielo.org.bo/scielo.php?pid=S2072-92942017000200008&script=sci_arttext&tlng=pt

GUERRERO, Fernanda Soto. ANÁLISIS DE LA PROBLEMÁTICA ASOCIADA CON LA BAJA CALIDAD DE DATOS EN LOS SISTEMAS DE INFORMACIÓN. 2014. Recuperado de https://core.ac.uk/download/pdf/148685230.pdf

HERNÁNDEZ ESCOBAR, Arturo, et al., 2018. Metodología de la investigación científica 3Ciencias. ISBN 9788494825705, disponible en: https://corladancash.com/wp-content/uploads/2020/01/Metodologia-de-la-inv-cientifica-Arturo-Andres-Hernandez-Escobar.pdf

HERNÁNDEZ-SAMPIERI, Roberto; MENDOZA, Christian. Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta. México. McGrawHill, 2018, Recuperado de http://repositorio.uasb.edu.bo:8080/handle/54000/1292.

INEI, Principales Resultados de la Encuesta Nacional de Empresas, recuperado de: https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1430/pdfs/libro.pdf

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2015. ISO/IEC 25012:2015-Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality disponible en: https://iso25000.com/index.php/normas-iso-25000/iso-25012.

JAYA, Izham; SIDI, Fatimah; ISHAK, Iskandar; AFFENDEY, Lilly; A. JABAR, Marzanah. 2017. A review of data quality research in achieving high data quality within organization. Journal of Theoretical and Applied Information Technology (JATIT). p.2647-2657. DOI: https://www.researchgate.net/publication/318490631_A_review_of_data_quality_research_in_achieving_high_data_quality_within_organization

KOI-AKROFI, Godfred; KOI-AKROFI, Joyce; AKWETEY, Henry. Understanding the characteristics, benefits and challenges of Agile IT Project Management: A literature based perspective. International Journal of Software Engineering & Applications (IJSEA), 2019, p. 25-44, Recuperado de https://arxiv.org/ftp/arxiv/papers/1910/1910.06218.pdf

LÓPEZ, Jaime. Web scraping. 2018. Disponible en https://www.academia.edu/35895308/Web_scraping

LUIZ, André; DA ROCHA, Iury; BRENDON, Jonathan; DA SILVA, Marina; BARUCKE, Guilherme. Scrum-Based Application for Agile Project Management. Journal of Software (JSW), 2019, p.106-113, Recuperado de http://www.jsoftware.us/vol15/421-T1035.pdf

MARTINEZ NUÑEZ, Antonio Federico. Automatización de web Scraping de los diarios de noticias para la empresa Isuri, San Martín de Porres, 2020, recuperado de https://repositorio.ucv.edu.pe/handle/20.500.12692/48352

MÉNDEZ MATAMOROS, Jorge Hernando, et al. Mejoramiento de calidad en conjuntos de datos abiertos basado en la aplicación de métricas de consistencia lógica. 2017. Disponible en http://repository.udistrital.edu.co/handle/11349/8032

MITCHELL, Ryan. Web scraping with Python: Collecting more data from the modern web. " O'Reilly Media, Inc.", 2018. Disponible en https://edu.anarcho-copy.org/Programming%20Languages/Python/Web%20Scraping%20with%20Python, %202nd%20Edition.pdf.

NANDWANI, Usha; MISHRA, Ritesh; PATIL, Amol; SIDDIQUI, Wasimudin. Data Analysis by Web Scraping using Python. International Journal for Research in Engineering Application & Management (IJREAM), 2021, p. 12-15, Recuperado de http://ijream.org/papers/IJREAMV07I02SJ003.pdf

PAD-RTM (2020), Encuesta de Transformación Digital, Recuperado de https://www.rtm.com.pe/wp-content/uploads/2021/01/Encuesta-TD-2020.pdf

PESADO, Patricia Mabel; ARROYO, Marcelo. XXV Congreso Argentino de Ciencias de la Computación-CACIC 2019: libro de actas. 2020. Disponible en http://sedici.unlp.edu.ar/handle/10915/90359

PRÍNCIPE QUIÑONES, Brady Marlon; MENDOZA RUIZ, Christiam Alberth. Automatización robótica de procesos en las conciliaciones bancarias de una empresa industrial. 2019, publicado en https://repositorio.upn.edu.pe/handle/11537/22495

RINCÓN RODRÍGUEZ, Magaly. Plan de gestión de calidad de datos para mejorar la oportunidad y pertinencia de la información de la oferta institucional en la Dirección de Apropiación del Ministerio TIC. 2019. Tesis Doctoral. Bogotá: Universidad Externado de Colombia, 2019., disponible en https://bdigital.uexternado.edu.co/handle/001/2451

RODRÍGUEZ GARCÍA, Diego, et al. Desarrollo RPA para monitoreo de calidad de datos y generación de alertas. 2020. Tesis Doctoral. Universidad EAFIT, Recuperado de https://repository.eafit.edu.co/handle/10784/17555

SALAZAR PINTO, Cecilia; DEL CASTILLO GALARZA, Santiago. Fundamentos básicos de estadística. 2017. Recuperado de http://www.dspace.uce.edu.ec/handle/25000/13720

SAURKAR, Anand V.; PATHARE, Kedar G .; GODE, Shweta A. An Overview On Web Scraping Techniques And Tools. International Journal on Future Revolution in Computer Science & Communication Engineering, 2018, DOI: http://www.ijfrcsce.org/index.php/ijfrcsce/article/view/1529

SCHWABER, Ken; SUTHERLAND, Jeff. The Scrum Guide (The Definitive Guide to Scrum: The Rules of the Game). 2017. Recuperado de https://scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-US.pdf.

SERNA CARVAJAL, Yennifer Vanessa. Automatización robótica de procesos (RPA). 2021, Recuperado de https://bibliotecadigital.udea.edu.co/handle/10495/19655

UCV, Resolución de Consejo Universitario, 2017, recuperado de https://www.ucv.edu.pe/datafiles/C%C3%93DIGO%20DE%20%C3%89TICA.pdf

VANCAUWENBERGH, Sadia. Data Quality Management. 2019. DOI: https://www.researchgate.net/publication/334201997_Data_Quality_Management

VEIGA, Allan Koch y col. A conceptual framework for quality assessment and management of biodiversity data. PloS One, 2017, vol. 12, no 6, pág. e0178731. Publicado en https://doi.org/10.1371/journal.pone.0178731

VELASCO CHÁVARRY, Rafael Hipólito. Sistema informático para el control de calidad de datos e información estadística en los establecimientos de CLAS Batanes. 2018, publicado en https://repositorio.ucv.edu.pe/handle/20.500.12692/43851

WILEY, John. Robotic Process Automation for Dummies. 2018. Recuperado de https://www.nice.com/rpa/assets/robotic_process_automation_for_dummies.pdf

WRIGHT, Tim; BOTT, Antony. Legal Issues in Robotic Process Automation. IDM, 2018, p.24-25. Recuperado de: https://idm.net.au/files/IDM-PDF-archive/2018-0203.pdf

YARLAGADDA, Ravi Teja. The RPA and AI Automation. International Journal of Creative Research Thoughts (IJCRT), ISSN, 2018, p. 2320-2882, recuperado de https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3798275

ZHAO, Bo. Web Scraping. 2017, p.1-3, DOI https://www.researchgate.net/publication/317177787_Web_Scraping

ANEXOS

Anexo 1: Matriz de consistencia.

Tabla 12: Tabla 12: Matriz de consistencia

PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES	DISEÑO DE LA INVESTIGACIÓN
PROBLEMA GENERAL: ¿De qué manera influye un sistema RPA utilizando técnicas de web scraping en la garantía de la calidad de datos en la empresa Konecta, Lima 2021? **PROBLEMAS ESPECÍFICOS:** ¿En qué medida un sistema RPA utilizando técnicas de web scraping influye en la consistencia de formato de datos para la garantía de la calidad de datos en la empresa Konecta, Lima 2021? ¿En qué medida un sistema RPA utilizando técnicas de web scraping influye en la frecuencia de actualización para la garantía de la calidad de datos en la empresa Konecta, Lima 2021?	**OBJETIVO GENERAL:** Determinar la influencia de un sistema RPA utilizando técnicas de web scraping para garantizar la calidad de datos en la empresa Konecta, Lima 2021. **OBJETIVOS ESPECÍFICOS:** Determinar la influencia de un sistema RPA utilizando técnicas de web scraping en la consistencia de formato de datos para garantizar la calidad de datos en la empresa Konecta, Lima 2021. Determinar la influencia de un sistema RPA utilizando técnicas de web scraping en la frecuencia de actualización para garantizar la calidad de datos en la empresa Konecta, Lima 2021.	**HIPÓTESIS GENERAL:** Un sistema RPA utilizando técnicas de web scraping garantiza la calidad de datos en la empresa Konecta, Lima 2021 **HIPÓTESIS ESPECÍFICAS:** Un sistema RPA utilizando técnicas de web scraping aumenta el grado de consistencia de formato de datos para la garantía de la calidad de datos en la empresa Konecta, Lima 2021. Un sistema RPA utilizando técnicas de web scraping incrementa el grado de frecuencia de actualización para la garantía de la calidad de datos en la empresa Konecta, Lima 2021.	**VARIABLE INDEPENDIENTE:** Sistema RPA utilizando técnicas de web scraping **VARIABLE DEPENDIENTE:** Calidad de datos	**ENFOQUE DE LA INVESTIGACIÓN:** Cuantitativo **TIPO DE INVESTIGACIÓN:** Aplicada **DISEÑO DE INVESTIGACIÓN:** Pre-experimental **POBLACIÓN Y MUESTRA:** 31 conjuntos de datos de atención al cliente **INSTRUMENTO:** Ficha de registro

Fuente: Elaboración propia de los autores

Anexo 2: Configuración definida de para los ítems de datos.

Para cumplir con la calidad de datos, el RPA tiene una tabla de configuración para los subprocesos de los 8 ítems que conforman cada conjunto de datos. Cada subproceso se encarga de recolectar datos distintos y poder exportarlos a un formato estructurado que le corresponde (CSV, XLSX, TXT).

Figura 10: Tabla de configuración para los ítems de datos.

Ítem	Frecuencia	Nombre	Formato Definido	Hora Definida
1	Diaria	BD_01	.xlsx	11:00 a. m.
2	Diaria	BD_02	.txt.gz	11:00 a. m.
3	Diaria	BD_03	.txt.gz	11:00 a. m.
4	Diaria	BD_04	.csv.gz	11:00 a. m.
5	Diaria	BD_05	.csv.gz	11:00 a. m.
6	Diaria	BD_06	.csv	11:00 a. m.
7	Diaria	BD_07	.csv	11:00 a. m.
8	Diaria	BD_08	.csv	11:00 a. m.

Fuente: Elaboración propia de los autores.

Anexo 3: Registro detallado del conjunto de datos generado por el RPA.

La siguiente tabla tiene el detalle del ítem de datos que recoge el RPA, de esta tabla podremos completar la ficha de registro. Podremos saber si cumplen formato o cumple frecuencia de actualización comparando con cada ítem del Anexo 2.

Figura 11: Tabla del detalle del conjunto de datos generado por el RPA.

Id	Ítem	Nombre	Peso	Actualización	Formato
20210801	1	BD_01	195 KB	14/09/2021 10:03	.xlsx
20210801	2	BD_02	196 KB	23/09/2021 09:20	.txt.gz
20210801	3	BD_03	76 MB	23/09/2021 09:21	.txt.gz
20210801	4	BD_04	4 MB	23/09/2021 09:21	.csv.gz
20210801	5	BD_05	321 MB	23/09/2021 09:23	.csv.gz
20210801	6	BD_06	187 MB	23/09/2021 09:23	.csv
20210801	7	BD_07	95 MB	23/09/2021 09:27	.csv
20210801	8	BD_08	15 MB	23/09/2021 09:30	.csv

Fuente: Elaboración propia de los autores.

Anexo 4: Ficha de registro pretest para el grado consistencia de formato de datos.

En la siguiente tabla se muestra la ficha de registro Pretest para el indicador grado de inconsistencia de datos.

Tabla 13: Ficha de registro pretest para el grado consistencia de formato de datos.

N° Ficha Registro		1		
Observador		Franklin Revilla - Anthony Herrera		
Institución donde se investiga		Konecta Perú		
Dirección				
Proceso de observación		**X = A/B**		
Id	**Fecha**	**A**	**B**	**X**
20210701	1/7/2021	7	8	0,88
20210702	2/7/2021	8	8	1,00
20210703	3/7/2021	7	8	0,88
20210704	4/7/2021	6	8	0,75
20210705	5/7/2021	8	8	1,00
20210706	6/7/2021	6	8	0,75
20210707	7/7/2021	5	8	0,63
20210708	8/7/2021	6	8	0,75
20210709	9/7/2021	6	8	0,75
20210710	10/7/2021	6	8	0,75
20210711	11/7/2021	7	8	0,88
20210712	12/7/2021	6	8	0,75
20210713	13/7/2021	6	8	0,75
20210714	14/7/2021	6	8	0,75
20210715	15/7/2021	6	8	0,75
20210716	16/7/2021	6	8	0,75
20210717	17/7/2021	6	8	0,75
20210718	18/7/2021	6	8	0,75
20210719	19/7/2021	6	8	0,75
20210720	20/7/2021	7	8	0,88
20210721	21/7/2021	8	8	0,63
20210722	22/7/2021	8	8	1,00
20210723	23/7/2021	6	8	0,75
20210724	24/7/2021	6	8	0,75
20210725	25/7/2021	5	8	1,00
20210726	26/7/2021	6	8	0,75
20210727	27/7/2021	6	8	0,75
20210728	28/7/2021	8	8	1,00
20210729	29/7/2021	7	8	0,88
20210730	30/7/2021	6	8	0,75
20210731	31/7/2021	7	8	0,88

Fuente: Elaboración propia de los autores.

Anexo 5: Ficha de registro pretest para el grado de frecuencia de actualización de datos.

En la siguiente tabla se muestra la ficha de registro Pretest para el indicador frecuencia de actualización de datos.

Tabla 14: Ficha de registro pretest para el grado de frecuencia de actualización de datos.

N° Ficha Registro		1		
Observador		Franklin Revilla - Anthony Herrera		
Institución donde se investiga		Konecta Perú		
Dirección				
Proceso de observación		**X = A/B**		
Id	**Fecha**	**A**	**B**	**X**
20210701	1/7/2021	7	8	0,88
20210702	2/7/2021	8	8	1,00
20210703	3/7/2021	7	8	0,88
20210704	4/7/2021	5	8	0,63
20210705	5/7/2021	8	8	1,00
20210706	6/7/2021	5	8	0,63
20210707	7/7/2021	5	8	0,63
20210708	8/7/2021	6	8	0,75
20210709	9/7/2021	6	8	0,75
20210710	10/7/2021	6	8	0,75
20210711	11/7/2021	7	8	0,88
20210712	12/7/2021	4	8	0,50
20210713	13/7/2021	6	8	0,75
20210714	14/7/2021	6	8	0,75
20210715	15/7/2021	6	8	0,75
20210716	16/7/2021	6	8	0,75
20210717	17/7/2021	6	8	0,75
20210718	18/7/2021	5	8	0,63
20210719	19/7/2021	6	8	0,75
20210720	20/7/2021	7	8	0,88
20210721	21/7/2021	8	8	0,50
20210722	22/7/2021	6	8	0,75
20210723	23/7/2021	5	8	0,63
20210724	24/7/2021	5	8	0,63
20210725	25/7/2021	4	8	1,00
20210726	26/7/2021	4	8	0,50
20210727	27/7/2021	5	8	0,63
20210728	28/7/2021	7	8	0,88
20210729	29/7/2021	6	8	0,75
20210730	30/7/2021	5	8	0,63
20210731	31/7/2021	4	8	0,50

Fuente: Elaboración propia de los autores.

Anexo 6: Ficha de registro postest para el grado de consistencia de datos.

En la siguiente tabla se muestra la ficha de registro Postest para el indicador grado de inconsistencia de datos.

Tabla 15: Ficha de registro postest para el grado de consistencia de datos.

N° Ficha Registro		1		
Observador		Franklin Revilla - Anthony Herrera		
Institución donde se investiga		Konecta Perú		
Dirección				
Proceso de observación		**X = A/B**		
Id	**Fecha**	**A**	**B**	**X**
20210801	1/8/2021	8	8	1,00
20210802	2/8/2021	8	8	1,00
20210803	3/8/2021	8	8	1,00
20210804	4/8/2021	8	8	1,00
20210805	5/8/2021	8	8	1,00
20210806	6/8/2021	8	8	1,00
20210807	7/8/2021	8	8	1,00
20210808	8/8/2021	8	8	1,00
20210809	9/8/2021	8	8	1,00
20210810	10/8/2021	8	8	1,00
20210811	11/8/2021	8	8	1,00
20210812	12/8/2021	8	8	1,00
20210813	13/8/2021	8	8	1,00
20210814	14/8/2021	8	8	1,00
20210815	15/8/2021	8	8	1,00
20210816	16/8/2021	8	8	1,00
20210817	17/8/2021	8	8	1,00
20210818	18/8/2021	8	8	1,00
20210819	19/8/2021	8	8	1,00
20210820	20/8/2021	8	8	1,00
20210821	21/8/2021	8	8	1,00
20210822	22/8/2021	8	8	1,00
20210823	23/8/2021	8	8	1,00
20210824	24/8/2021	8	8	1,00
20210825	25/8/2021	8	8	1,00
20210826	26/8/2021	8	8	1,00
20210827	27/8/2021	8	8	1,00
20210828	28/8/2021	8	8	1,00
20210829	29/8/2021	8	8	1,00
20210830	30/8/2021	8	8	1,00
20210831	31/8/2021	8	8	1,00

Fuente: Elaboración propia de los autores

Anexo 7: Ficha de registro postest para el grado de frecuencia de actualización de datos.

En la siguiente tabla se muestra la ficha de registro postest para el indicador frecuencia de actualización de datos.

Tabla 16: Ficha de registro postest para el grado de frecuencia de actualización de datos.

N° Ficha Registro		1		
Observador		Franklin Revilla - Anthony Herrera		
Institución donde se investiga		Konecta Perú		
Dirección				
Proceso de observación		**X = A/B**		
Id	**Fecha**	**A**	**B**	**X**
20210801	1/8/2021	8	8	1,00
20210802	2/8/2021	8	8	1,00
20210803	3/8/2021	8	8	1,00
20210804	4/8/2021	8	8	1,00
20210805	5/8/2021	8	8	1,00
20210806	6/8/2021	8	8	1,00
20210807	7/8/2021	8	8	1,00
20210808	8/8/2021	8	8	1,00
20210809	9/8/2021	8	8	1,00
20210810	10/8/2021	8	8	1,00
20210811	11/8/2021	8	8	1,00
20210812	12/8/2021	8	8	1,00
20210813	13/8/2021	8	8	1,00
20210814	14/8/2021	8	8	1,00
20210815	15/8/2021	8	8	1,00
20210816	16/8/2021	8	8	1,00
20210817	17/8/2021	8	8	1,00
20210818	18/8/2021	8	8	1,00
20210819	19/8/2021	8	8	1,00
20210820	20/8/2021	8	8	1,00
20210821	21/8/2021	8	8	1,00
20210822	22/8/2021	8	8	1,00
20210823	23/8/2021	8	8	1,00
20210824	24/8/2021	8	8	1,00
20210825	25/8/2021	8	8	1,00
20210826	26/8/2021	8	8	1,00
20210827	27/8/2021	8	8	1,00
20210828	28/8/2021	7	8	0,88
20210829	29/8/2021	8	8	1,00
20210830	30/8/2021	8	8	1,00
20210831	31/8/2021	7	8	0,88

Fuente: Elaboración propia de los autores.

Anexo 8: Resultados de la prueba TEST-RETEST

Tabla 17: Ficha de frecuencia de actualización - Test

N° Ficha Registro		1		
Observador		Franklin Revilla - Anthony Herrera		
Institucion donde se investiga		Konecta Perú		
Dirección				
Proceso de observación		Frecuencia de Actualización		X = A/B
Id	Fecha	A	B	X
20210706	6/8/2021	5	8	0,63
20210707	7/8/2021	5	8	0,63
20210708	8/8/2021	6	8	0,75
20210709	9/8/2021	6	8	0,75
20210710	10/8/2021	6	8	0,75
20210711	11/8/2021	7	8	0,88
20210712	12/8/2021	4	8	0,50

Fuente: Elaboración propia de los autores.

Tabla 18: Ficha de frecuencia de actualización - Retest.

N° Ficha Registro		1		
Observador		Franklin Revilla - Anthony Herrera		
Institucion donde se investiga		Konecta Perú		
Dirección				
Proceso de observación		Frecuencia de Actualización		X = A/B
Id	Fecha	A	B	X
20210715	15/8/2021	6	8	0,75
20210716	16/8/2021	6	8	0,75
20210717	17/8/2021	6	8	0,75
20210718	18/8/2021	5	8	0,63
20210719	19/8/2021	6	8	0,75
20210720	20/8/2021	7	8	0,88
20210721	21/8/2021	8	8	0,50

Fuente: Elaboración propia de los autores.

Tabla 19: Prueba de Normalidad para la prueba Test-Retest.

	Shapiro-Wilk		
	Estadístico	gl	Sig.
FDA_test	0.937	7	0.609
FDA_retest	0.869	7	0.183

Fuente: Elaboración propia de los autores.

Anexo 9: Cálculo al detalle de la confiabilidad de los instrumentos.

Coeficiente de correlación de Pearson (Los datos de la prueba Test-Retest tienen una distribución normal).

Tabla 20: Tabla del coeficiente de correlación de Pearson para la prueba de Test-Retest.

		FDA_test	FDA_retest
FDA_test	Correlación de Pearson	1,000	,744
	Sig. (bilateral)		,055
	N	7	7
FDA_retest	Correlación de Pearson	,744	1,000
	Sig. (bilateral)	,055	
	N	7	7

Fuente: Elaboración propia de los autores.

El valor del Coeficiente de correlación de Pearson es 0.744, el cual es claramente mayor que 0.7, por lo tanto, queda determinado que el instrumento es confiable.

Anexo 10: Evaluación de juicio de expertos para determinar la metodología de desarrollo y la validez del instrumento.

TABLA DE EVALUACIÓN DE EXPERTOS

Apellidos y Nombres de Experto: MONTOYA NEGRILLO, DANY

Título y/o Grado:

Ph. D. () Doctor () Magister (X) Ingeniero () Otros: …….

Universidad que labora:

Fecha: 15/06/2021

TÍTULO DE PROYECTO

"SISTEMA RPA UTILIZANDO TÉCNICAS DE WEB SCRAPING PARA GARANTIZAR LA CALIDAD DE DATOS EN LA EMPRESA KONECTA, LIMA 2021"

Tabla de evaluación de expertos para la elección del marco de trabajo

Mediante la tabla de evaluación de expertos, usted tiene la facultad de calificar los marcos de trabajo involucrados, mediante una serie de preguntas marcando un valor en las columnas.

ITEM	CRITERIOS	MARCO DE TRABAJO		
		XP	AUP	SCRUM
1	Fomenta una mejor comunicación entre el cliente y los desarrolladores	1	2	3
2	Es el más destacado de los procesos agiles de desarrollo software, ocasiona eficiencia en el proceso de planificación y pruebas.	2	1	3
3	Brinda satisfacción al programador simplificando el diseño para agilizar el desarrollo y facilitar el mantenimiento.	2	2	3
4	Capaz de facilitar y adaptarse a los cambios de requisitos.	2	2	2
5	El cliente se involucra con el proyecto.	2	2	3
6	Pruebas continuas, frecuentemente repetidas y automatizadas, incluyendo pruebas de regresión en donde el cliente tiene control sobre las prioridades	3	3	3
TOTAL		12	12	17

Evaluar con la siguiente calificación:

1. Malo 2. Regular 3. Bueno

Sugerencias:

_____

_____

_____

Firma del Experto

VALIDACION DE INSTRUMENTOS – EVALUACION DE EXPERTOS

Apellidos y nombres del experto: MONTOYA NEGRILLO, DANY

Título y/o Grado: Magister en Ingeniería de Sistemas

Fecha: /……/……

Institución que labora: Universidad Cesar Vallejo

Título del proyecto Investigación: SISTEMA RPA UTILIZANDO TÉCNICAS DE WEB SCRAPING PARA GARANTIZAR LA CALIDAD DE DATOS, LIMA 2021

Indicador: CONSISTENCIA DE FORMATO DE DATOS

Mediante la tabla de evaluación de expertos, usted tiene la facultad de evaluar a cada uno de los criterios indicando el valor porcentual.

Indicadores	**CRITERIOS**	**Deficiente 0% - 24%**	**Regular 25% - 49%**	**Bueno 50% - 60%**	**Muy Bueno 61% - 74%**	**Excelente 75% - 100%**
Claridad	El instrumento d recolección de datos se relaciona con la variable de investigación					95
Organización	Sera accesible a la población sujeto de estudio					95
Metodología	El instrumento de recolección de datos facilitara el logro de los objetivos de la investigación					95
Objetividad	El instrumento de recolección de datos menciona las variables de la investigación					95
Pertinencia	El diseño del instrumento de medición facilitara el análisis y procesamiento de datos					95
	TOTAL					95%

Resultado: ……

Aplicabilidad: El instrumento puede ser aplicado (X) El instrumento debe ser mejorado ( )

Observaciones:

……
……

Firma del Experto

TABLA DE EVALUACIÓN DE EXPERTOS

Apellidos y Nombres de Experto: HILARIO FALCON, FRANCISCO MANUEL

Título y/o Grado:

Ph. D. () Doctor (X) Magister () Ingeniero () Otros: …….

Universidad que labora:

Fecha:

TÍTULO DE PROYECTO

"SISTEMA RPA UTILIZANDO TÉCNICAS DE WEB SCRAPING PARA GARANTIZAR LA CALIDAD DE DATOS, LIMA 2021"

Tabla de evaluación de expertos para la elección del marco de trabajo

Mediante la tabla de evaluación de expertos, usted tiene la facultad de calificar los marcos de trabajo involucrados, mediante una serie de preguntas marcando un valor en las columnas.

ITEM	CRITERIOS	MARCO DE TRABAJO		
		XP	AUP	SCRUM
1	Fomenta una mejor comunicación entre el cliente y los desarrolladores	3	3	3
2	Es el más destacado de los procesos agiles de desarrollo software, ocasiona eficiencia en el proceso de planificación y pruebas.	3	3	3
3	Brinda satisfacción al programador simplificando el diseño para agilizar el desarrollo y facilitar el mantenimiento.	3	2	3
4	Capaz de facilitar y adaptarse a los cambios de requisitos.	3	2	3
5	El cliente se involucra con el proyecto.	3	3	3
6	Pruebas continuas, frecuentemente repetidas y automatizadas, incluyendo pruebas de regresión en donde el cliente tiene control sobre las prioridades	2	3	3
TOTAL		17	17	18

Evaluar con la siguiente calificación:

1. Malo 2. Regular 3. Bueno

Sugerencias:

_____

Dr. Francisco Manuel Hilario Falcón

DNI: 10132075

VALIDACION DE INSTRUMENTOS – EVALUACION DE EXPERTOS

Apellidos y nombres del experto: HILARIO FALCON, FRANCISCO MANUEL

Título y/o Grado: DOCTOR EN INGENIERIA DE SISTEMAS

Fecha: ……/……/……

Institución que labora: UNIVERSIDAD CESAR VALLEJO

Título del proyecto Investigación: SISTEMA RPA UTILIZANDO TÉCNICAS DE WEB SCRAPING PARA GARANTIZAR LA CALIDAD DE DATOS, LIMA 2021

Indicador: CONSISTENCIA DE FORMATO DE DATOS

Mediante la tabla de evaluación de expertos, usted tiene la facultad de evaluar a cada uno de los criterios indicando el valor porcentual.

Indicadores	**CRITERIOS**	**Deficiente 0% - 24%**	**Regular 25% - 49%**	**Bueno 50% - 60%**	**Muy Bueno 61% - 74%**	**Excelente 75% - 100%**
Claridad	El instrumento d recolección de datos se relaciona con la variable de investigación					95
Organización	Sera accesible a la población sujeto de estudio					95
Metodología	El instrumento de recolección de datos facilitara el logro de los objetivos de la investigación					95
Objetividad	El instrumento de recolección de datos menciona las variables de la investigación					95
Pertinencia	El diseño del instrumento de medición facilitara el análisis y procesamiento de datos					95
	TOTAL					95%

Resultado: 95%……

Aplicabilidad: El instrumento puede ser aplicado (X) El instrumento debe ser mejorado ( )

Observaciones:

……
……

Dr. Francisco Manuel Hilario Falcón

DNI: 10132075

TABLA DE EVALUACIÓN DE EXPERTOS

Apellidos y Nombres de Experto: PETRLIK AZABACHE, IVAN CARLO

Título y/o Grado:

Ph. D. () Doctor (X) Magister () Ingeniero () Otros: …….

Universidad que labora:

Fecha:

TÍTULO DE PROYECTO

"SISTEMA RPA UTILIZANDO TÉCNICAS DE WEB SCRAPING PARA GARANTIZAR LA CALIDAD DE DATOS"

Tabla de evaluación de expertos para la elección del marco de trabajo

Mediante la tabla de evaluación de expertos, usted tiene la facultad de calificar los marcos de trabajo involucrados, mediante una serie de preguntas marcando un valor en las columnas.

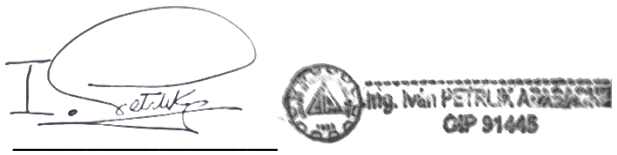
ITEM	CRITERIOS	MARCO DE TRABAJO		
		XP	AUP	SCRUM
1	Fomenta una mejor comunicación entre el cliente y los desarrolladores	3	3	3
2	Es el más destacado de los procesos agiles de desarrollo software, ocasiona eficiencia en el proceso de planificación y pruebas.	3	2	3
3	Brinda satisfacción al programador simplificando el diseño para agilizar el desarrollo y facilitar el mantenimiento.	3	2	3
4	Capaz de facilitar y adaptarse a los cambios de requisitos.	2	2	3
5	El cliente se involucra con el proyecto.	2	2	3
6	Pruebas continuas, frecuentemente repetidas y automatizadas, incluyendo pruebas de regresión en donde el cliente tiene control sobre las prioridades	3	2	3
TOTAL		16	13	18

Evaluar con la siguiente calificación:

1. Malo 2. Regular 3. Bueno

Sugerencias:

_____



Firma del Experto
Dni : 10140461

VALIDACION DE INSTRUMENTOS – EVALUACION DE EXPERTOS

Apellidos y nombres del experto: PETRLIK AZABACHE, IVAN CARLO

Título y/o Grado: Doctor en Ingeniería de Sistemas e Investigador RENACYT - CONCYTEC

Fecha: ……/……/……

Institución que labora: UNIVERSIDAD CESAR VALLEJO

Título del proyecto Investigación: SISTEMA RPA UTILIZANDO TÉCNICAS DE WEB SCRAPING PARA GARANTIZAR LA CALIDAD DE DATOS EN LA EMPRESA KONECTA, LIMA 2021

Indicador: CONSISTENCIA DE FORMATO DE DATOS

Mediante la tabla de evaluación de expertos, usted tiene la facultad de evaluar a cada uno de los criterios indicando el valor porcentual.

Indicadores	**CRITERIOS**	**Deficiente 0% - 24%**	**Regular 25% - 49%**	**Bueno 50% - 60%**	**Muy Bueno 61% - 74%**	**Excelente 75% - 100%**
Claridad	El instrumento de recolección de datos se relaciona con la variable de investigación					90
Organización	Sera accesible a la población sujeto de estudio					76
Metodología	El instrumento de recolección de datos facilitara el logro de los objetivos de la investigación					98
Objetividad	El instrumento de recolección de datos menciona las variables de la investigación					82
Pertinencia	El diseño del instrumento de medición facilitara el análisis y procesamiento de datos					80
	TOTAL					85.2

Resultado: 85.2 %

Aplicabilidad: El instrumento puede ser aplicado (X) El instrumento debe ser mejorado ( )

Observaciones:

…… ……
……

Firma del Experto

Dni : 10140461

VALIDACION DE INSTRUMENTOS – EVALUACION DE EXPERTOS

Apellidos y nombres del experto: PETRLIK AZABACHE, IVAN CARLO

Título y/o Grado: Doctor en Ingeniería de Sistemas e Investigador RENACYT - CONCYTEC

Fecha: ……/……/……

Institución que labora: Doctor en Ingeniería de Sistemas e Investigador RENACYT - CONCYTEC

Título del proyecto Investigación: SISTEMA RPA UTILIZANDO TÉCNICAS DE WEB SCRAPING PARA GARANTIZAR LA CALIDAD DE DATOS EN LA EMPRESA KONECTA, LIMA 2021

Indicador: FRECUENCIA DE ACTUALIZACIÓN

Mediante la tabla de evaluación de expertos, usted tiene la facultad de evaluar a cada uno de los criterios indicando el valor porcentual.

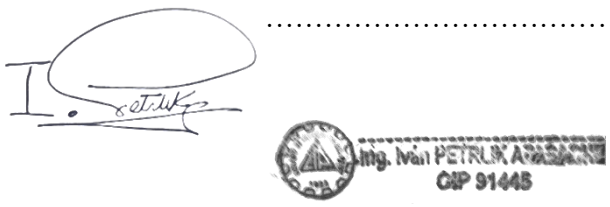
Indicadores	**CRITERIOS**	**Deficiente 0% - 24%**	**Regular 25% - 49%**	**Bueno 50% - 60%**	**Muy Bueno 61% - 74%**	**Excelente 75% - 100%**
Claridad	El instrumento d recolección de datos se relaciona con la variable de investigación					80
Organización	Sera accesible a la población sujeto de estudio					85
Metodología	El instrumento de recolección de datos facilitara el logro de los objetivos de la investigación					90
Objetividad	El instrumento de recolección de datos menciona las variables de la investigación					86
Pertinencia	El diseño del instrumento de medición facilitara el análisis y procesamiento de datos					80
	TOTAL					84.2

Resultado: 84.2 %

Aplicabilidad: El instrumento puede ser aplicado (X) El instrumento debe ser mejorado ( )

Observaciones:

…… ……
……



Firma del Experto
Dni: 10140461

Anexo 11: Tabla de características de calidad de datos.

Figura 12: Tabla de características para la calidad de los datos.

CARACTERÍSTICA	INHERENTE	DEPENDIENTE DEL SISTEMA
Exactitud	X	
Completitud	X	
Consistencia	X	
Credibilidad	X	
Actualidad	X	
Accesibilidad	X	X
Conformidad	X	X
Confidencialidad	X	X
Eficiencia	X	X
Precisión	X	X
Trazabilidad	X	X
Comprensibilidad	X	X
Disponibilidad		X
Portabilidad		X
Recuperabilidad		X

Fuente: XXV Congreso Argentino de Ciencias de la Computación CACIC 2019 (2020).

Anexo 12: Documento que certifica el desarrollo de la tesis en la empresa, sellado.

AUTORIZACIÓN PARA REALIZAR LA DIFUSIÓN DE RESULTADOS DE LA INVESTIGACIÓN

Por medio del presente documento, yo Anthony Jean Claude Simon Altamirano, identificado con DNI N° 70405567 y Responsable de Business Intelligence del área Business Intelligence y Advanced Analytics del departamento de Tecnología de KONECTA PERÚ. Autorizo a Franklin Revilla Vergaray con DNI N° 73616052 y Luis Anthony Herrera Portella con DNI N° 46763914 a realizar la investigación titulada "Sistema RPA utilizando técnicas de web scraping para garantizar la calidad de datos en la empresa Konecta, Lima 2021" y difundir los resultados utilizando el nombre KONECTA PERÚ.

Lima, 23 de agosto del 2021.

Firma

Anthony Jean Claude Simon Altamirano

DNI N° 70405567

Cargo: Responsable de Business Intelligence

KONECTA PERÚ

Anexo 13: Documentación del sistema.

Anexo 14.1. Metodología de desarrollo.

1.- Metodología ágil (FRAMEWORK SCRUM)

Para el desarrollo del Sistema RPA se utilizará SCRUM, ya que nos permite ser flexibles y completar el desarrollo en poco tiempo, también que ayuda que el cliente en este caso, participe activamente del desarrollo a través de reuniones y entrevistas constantes. Además, al trabajar con sprints los entregables se podrán revisar de manera temprana.

Los roles son los siguientes:

- Scrum Master, gestor de los procesos y recursos utilizados en el desarrollo del producto.
- Product Owner o dueño del producto, encargado de recibir los requerimientos del cliente y administrador del producto.
- Equipo de desarrollo, conjunto de personas encargadas del desarrollo del producto.

Para el proyecto, utilizaremos los siguientes artefactos:

- Pila de producto o Product Backlog
- Pila de Sprint o Sprint Backlog

De igual forma utilizaremos las historias de usuario siendo estas una descripción breve, informal y en lenguaje sencillo de lo que un usuario quiere hacer dentro de un producto de software para obtener algo que le resulte valioso.

Se consideró para el desarrollo 4 sprints considerando en cada uno el desarrollo de cada historia de usuario, pruebas de QA para detectar algún defecto y la corrección de estos, finalmente las pruebas de aceptación donde el usuario da la conformidad del desarrollo realizado.

2.- Definición del Alcance.

El sistema RPA utilizando técnicas de web scraping está desarrollado y configurado para extraer datos únicamente de la web definida para este proyecto. Debido a que la web donde están alojado los datos es privada y solo se abre en una red especifica. Además de que los diferentes ítems de datos a extraer presentan características distintas como el formato, peso y numero de columnas.

PRODUCT BACKLOG:

CÓDIGO	HISTORIA DE USUARIO	DURACIÓN	PRIORIDAD
HU01	Preparar componentes para el desarrollo	7 días	Alta
HU02	Crear tabla de configuración	4 días	Media
HU03	Lógica para la Fecha de datos a extraer	3 días	Baja
HU04	Crear conexión a BD SQL Server	8 días	Alta
HU05	Conexión a web driver	4 días	Media
HU06	Simular inicio de sesión en web	2 días	Baja
HU07	Identificar xpath de los elementos	4 días	Media
HU08	Recuperar propiedades de la BD	7 días	Alta
HU09	Módulo de envió correo	3 días	Baja
HU10	Manejo de Excepciones	4 días	Media
HU11	Insertar propiedades en SQL	10 días	Alta

SPRINT 1:

HISTORIAS DE USUARIO

HU01	**HISTORIA DE USUARIO**
TÍTULO: PREPARAR COMPONENTES PARA EL DESARROLLO	
DESCRIPCIÓN: Las herramientas, entorno de trabajo deben estar listos.	
PRIORIDAD: Alta	**ESFUERZO:** 7 días
CRITERIO DE ACEPTACIÓN	
- Acceso a la web privada. - SQL Server 2017, python, web Driver deben estar instalados. - Librerías pandas, selenium, sqlAlchemy, pretty_html_table.	

HU02	**HISTORIA DE USUARIO**
TÍTULO: CREAR TABLA DE CONFIGURACIÓN	
DESCRIPCIÓN: Tabla que contiene nombre de ítem de datos, fecha, formato y ruta donde será guardado.	
PRIORIDAD: Media	**ESFUERZO:** 4 días
CRITERIO DE ACEPTACIÓN	
- Las fechas deben ser del día anterior (Hoy - 1). - Tener el formato y la hora definida para los ítems de datos.	

HU03	**HISTORIA DE USUARIO**
TÍTULO: LÓGICA PARA LA FECHA DE DATOS A EXTRAER	
DESCRIPCIÓN: Crear un procedimiento almacenado para realizar el cruce de las fechas de la tabla configuración con la tabla general, para validar si el ítem de dato ya fue cargado.	
PRIORIDAD: Baja	**ESFUERZO:** 3 días
CRITERIO DE ACEPTACIÓN:	
- El cruce es por cada Id de ítem y por fecha. - El resultado debe ser registros únicos (sin duplicidad).	

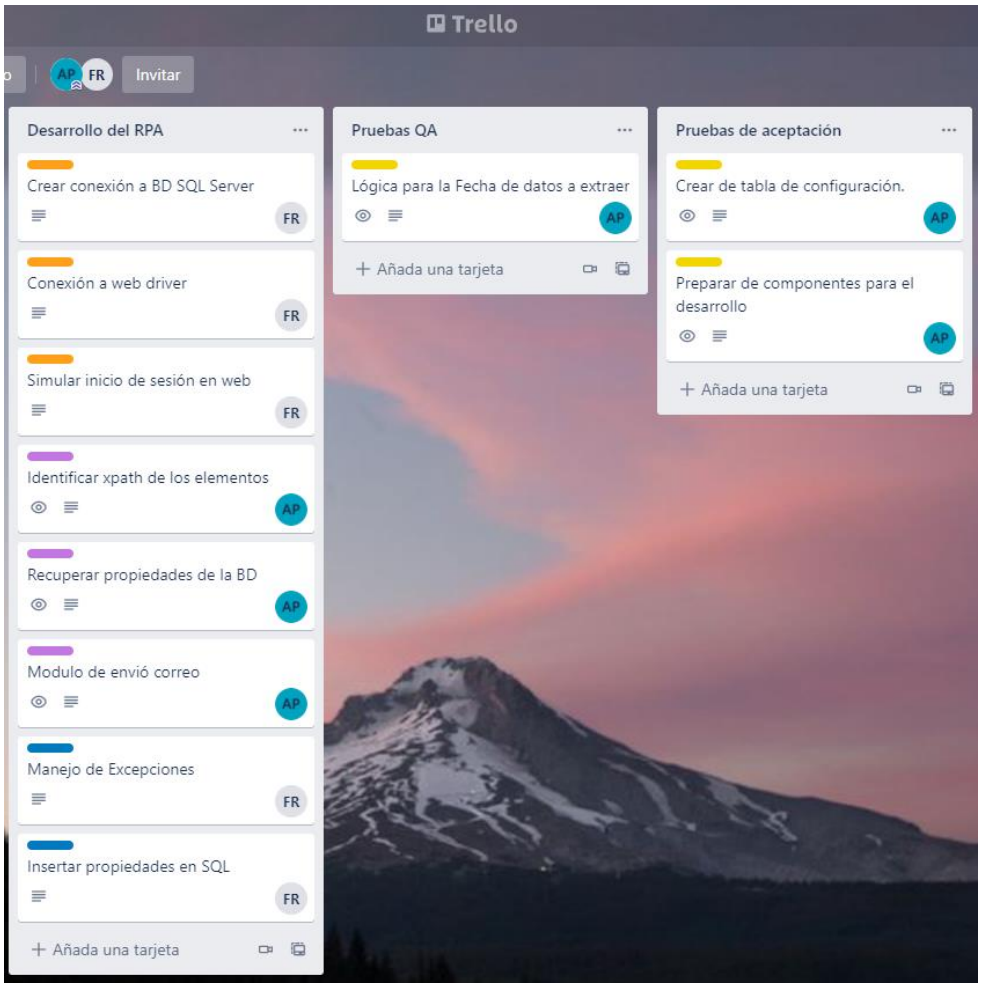
SPRINT BACKLOG

CÓDIGO	**HISTORIA DE USUARIO**	**ACTIVIDADES**
HU01	Preparar componentes para el desarrollo	- Acceso a la web privada. - SQL Server 2017, python, web Driver. - Librerías externas de python.
HU02	Crear tabla de configuración	- Crear tabla en SQL Server. - Insertar datos de configuración a la tabla.
HU03	Lógica para la Fecha de datos a extraer	- Cruce con tabla general. - Validar resultados únicos.

SPRINT REVIEW

CÓDIGO	HISTORIA DE USUARIO	CUMPLIO
HU01	Preparar componentes para el desarrollo	SI
HU02	Crear tabla de configuración	SI
HU03	Lógica para la Fecha de datos a extraer	SI

Figura 13: Tablero Kanban del Sprint 1.



Fuente: Elaboración propia de los autores.

SPRINT 2:

HISTORIAS DE USUARIO

HU04	**HISTORIA DE USUARIO**
TÍTULO: CREAR CONEXIÓN A BD SQL SERVER	
DESCRIPCIÓN: Cadena de conexión de python a SQL server usando sqlAlchemy.	
PRIORIDAD: Alta	**ESFUERZO:** 8 días
CRITERIO DE ACEPTACIÓN	
- Consulta de datos de SQL server. - Ejecución de procedimiento almacenado.	

HU05	**HISTORIA DE USUARIO**
TÍTULO: CONEXIÓN A WEB DRIVER	
DESCRIPCIÓN: Conectar a la web y levantar una pestaña de chrome con el URL de la página.	
PRIORIDAD: Media	**ESFUERZO:** 4 días
CRITERIO DE ACEPTACIÓN	
- La apertura de la web es correcta y rápida. - La configuración de descarga debe ser la adecuada para cada ítem.	

HU06	HISTORIA DE USUARIO
TÍTULO: SIMULAR INICIO DE SESIÓN EN WEB	
DESCRIPCIÓN: Mediante python y selenium simular la navegación por la web tal como si lo realizara una persona para extraer los datos.	
PRIORIDAD: Baja	**ESFUERZO:** 2 días
CRITERIO DE ACEPTACIÓN:	
- Ingresa usuario, contraseña y acepta términos. - Los datos para el acceso son recuperados de un Json.	

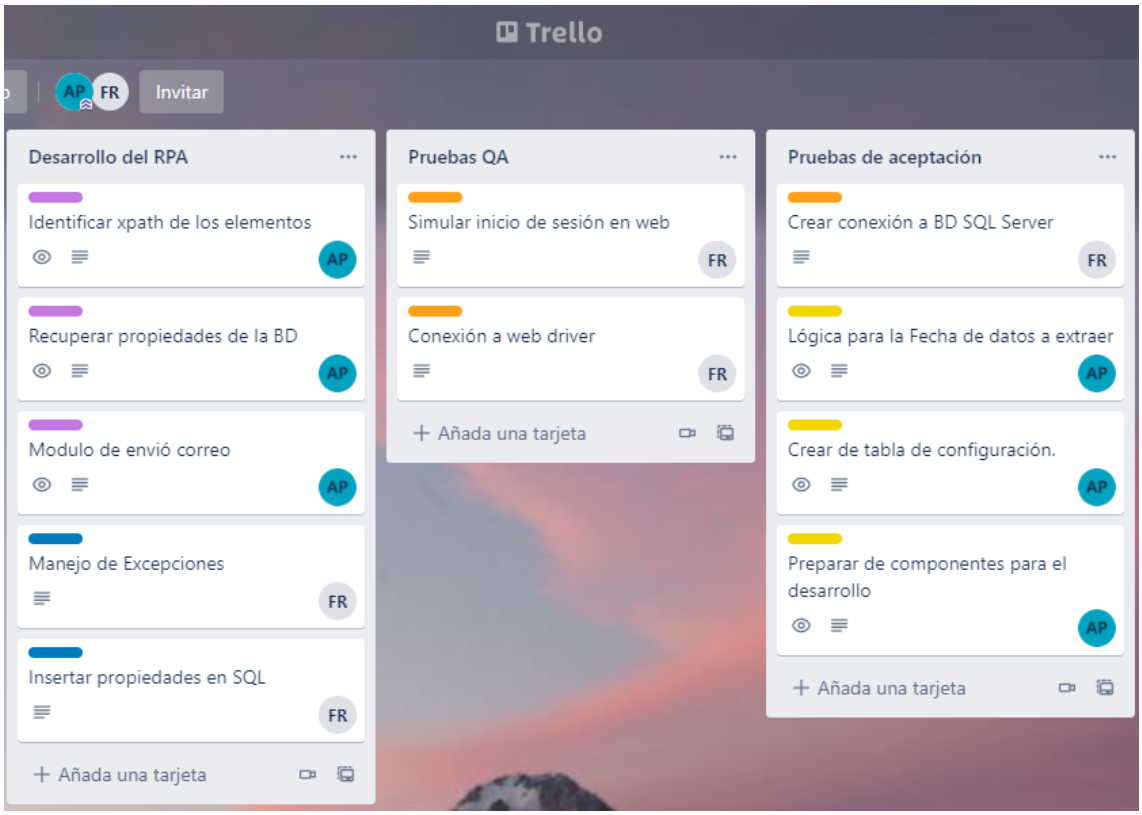
SPRINT BACKLOG

CÓDIGO	HISTORIA DE USUARIO	ACTIVIDADES
HU04	Crear conexión a BD SQL Server	- Conexión correcta a la BD. - Utilizar procedimientos almacenados de SQL.
HU05	Conexión a web driver	- Instalación de web diver Chrome 94.0.4606.61. - Configuración del Web diver.
HU06	Simular inicio de sesión en web	- Ingresar usuario, contraseña. - Aceptar condiciones. - Iniciar sesión.

SPRINT REVIEW

CÓDIGO	HISTORIA DE USUARIO	CUMPLIO
HU04	Crear conexión a BD SQL Server	SI
HU05	Conexión a web driver	SI
HU06	Simular inicio de sesión en web	SI

Figura 14: Tablero Kanban del Sprint 2.



Fuente: Elaboración propia de los autores.

SPRINT 3:

HISTORIAS DE USUARIO

HU07	**HISTORIA DE USUARIO**
TÍTULO: IDENTIFICAR XPATH DE LOS ELEMENTOS	
DESCRIPCIÓN: Para navegar en la web es necesario usar el Xpath para dar clic sobre botones o listas desplegables.	
PRIORIDAD: Media	**ESFUERZO:** 4 días
CRITERIO DE ACEPTACIÓN	
- Para identificar los ítems de datos se prioriza el uso de IDs de los elementos dentro del html. - Si se usa xpath, no tener errores al interactuar con estos elementos.	

HU08	**HISTORIA DE USUARIO**
TÍTULO: RECUPERAR PROPIEDADES DE LA BASE DE DATOS	
DESCRIPCIÓN: Se debe obtener el nombre, formato, fecha y hora de extracción de los datos.	
PRIORIDAD: Alta	**ESFUERZO:** 7 días
CRITERIO DE ACEPTACIÓN	
- Los campos mencionados en la descripción deben estar completos. - Recuperar propiedades de los 8 ítems de datos.	

HU09	HISTORIA DE USUARIO
TÍTULO: MÓDULO DE ENVÍO CORREO	
DESCRIPCIÓN: Módulo para enviar alertas de los eventos que tenga el sistema.	
PRIORIDAD: Baja	**ESFUERZO:** 3 días
CRITERIO DE ACEPTACIÓN:	
- Tener como cuerpo debe recibir un parámetro de tipo tabla. - Todas las propiedades de cada ítem de datos.	

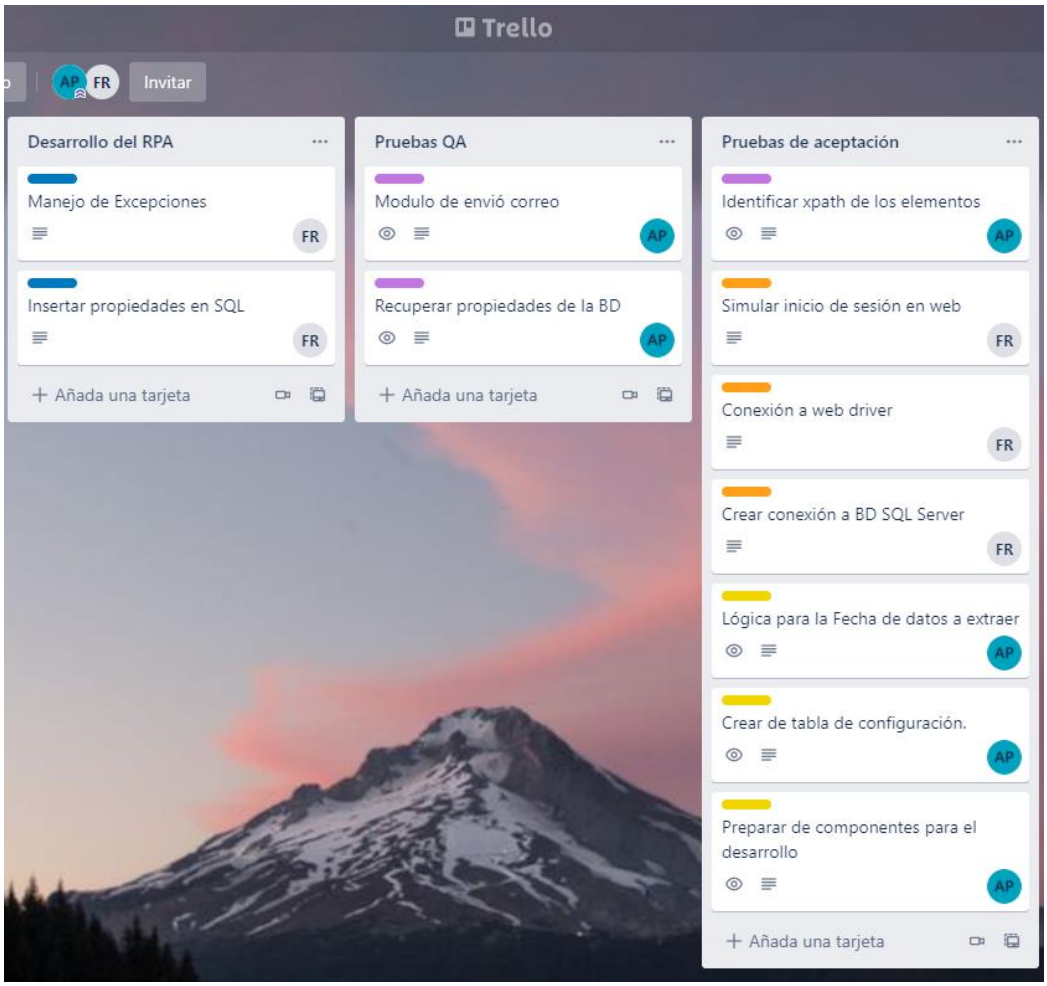
SPRINT BACKLOG

CÓDIGO	HISTORIA DE USUARIO	ACTIVIDADES
HU07	Identificar xpath de los elementos	- Para identificar los ítems de datos se prioriza el uso de IDs de los elementos del html
HU08	Recuperar propiedades de la BD	- Obtener el nombre, formato, fecha y hora de extracción de los ítems de datos
HU09	Módulo de envió correo	- El cuerpo del correo debe recibir un parámetro de tipo tabla. - Enviar las propiedades de cada ítem de datos.

SPRINT REVIEW

CÓDIGO	HISTORIA DE USUARIO	CUMPLIO
HU07	Identificar xpath de los elementos	SI
HU08	Recuperar propiedades de la BD	SI
HU09	Módulo de envió correo	SI

Figura 15: Tablero Kanban del Sprint 3.



Fuente: Elaboración propia de los autores.

SPRINT 4:

HISTORIAS DE USUARIO

HU10	**HISTORIA DE USUARIO**
TÍTULO: MANEJO DE EXCEPCIONES	
DESCRIPCIÓN: Los errores que tenga el sistema deben ser identificados.	
PRIORIDAD: Media	**ESFUERZO:** 4 días
CRITERIO DE ACEPTACIÓN	
- Detalla el tipo de error. - Detallar el paquete el error.	

HU11	**HISTORIA DE USUARIO**
TÍTULO: INSERTAR PROPIEDADES EN SQL	
DESCRIPCIÓN: Las propiedades obtenidas de los ítems de datos deben ser almacenados en una tabla de SQL Server.	
PRIORIDAD: Alta	**ESFUERZO:** 10 días
CRITERIO DE ACEPTACIÓN	
- No permitir registros duplicados. - Tener un ID de relación con tabla de configuración.	

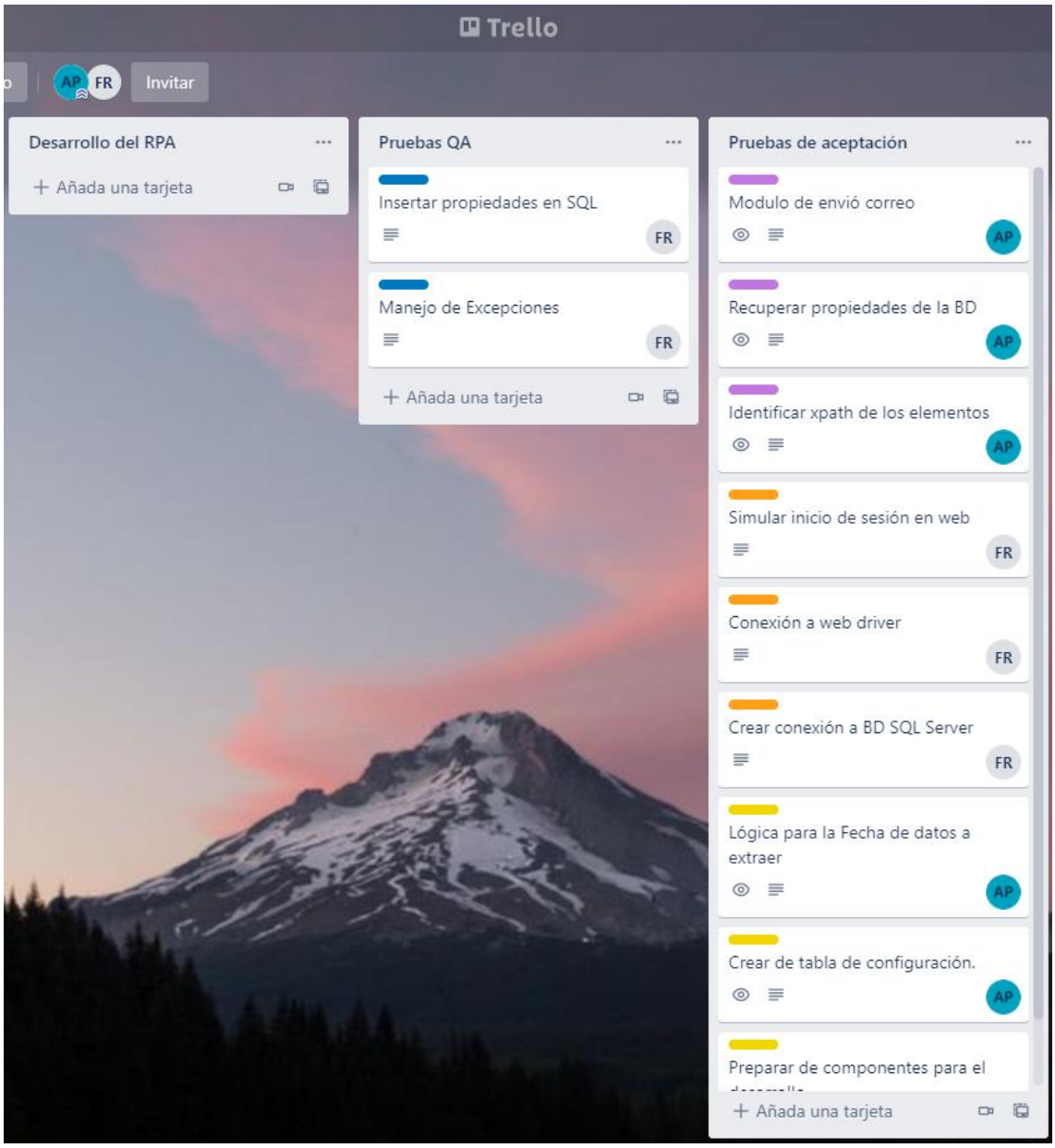
SPRINT BACKLOG

CÓDIGO	HISTORIA DE USUARIO	ACTIVIDADES
HU10	Manejo de Excepciones	- Identificar el tipo de error. - Detallar el paquete el error.
HU11	Insertar propiedades en SQL	- Insertar datos a SQL Server. - Tener un ID de relación con tabla de configuración.

SPRINT REVIEW

CÓDIGO	HISTORIA DE USUARIO	CUMPLIO
HU10	Manejo de Excepciones	SI
HU11	Insertar propiedades en SQL	SI

Figura 16:Tablero Kanban del Sprint 4.



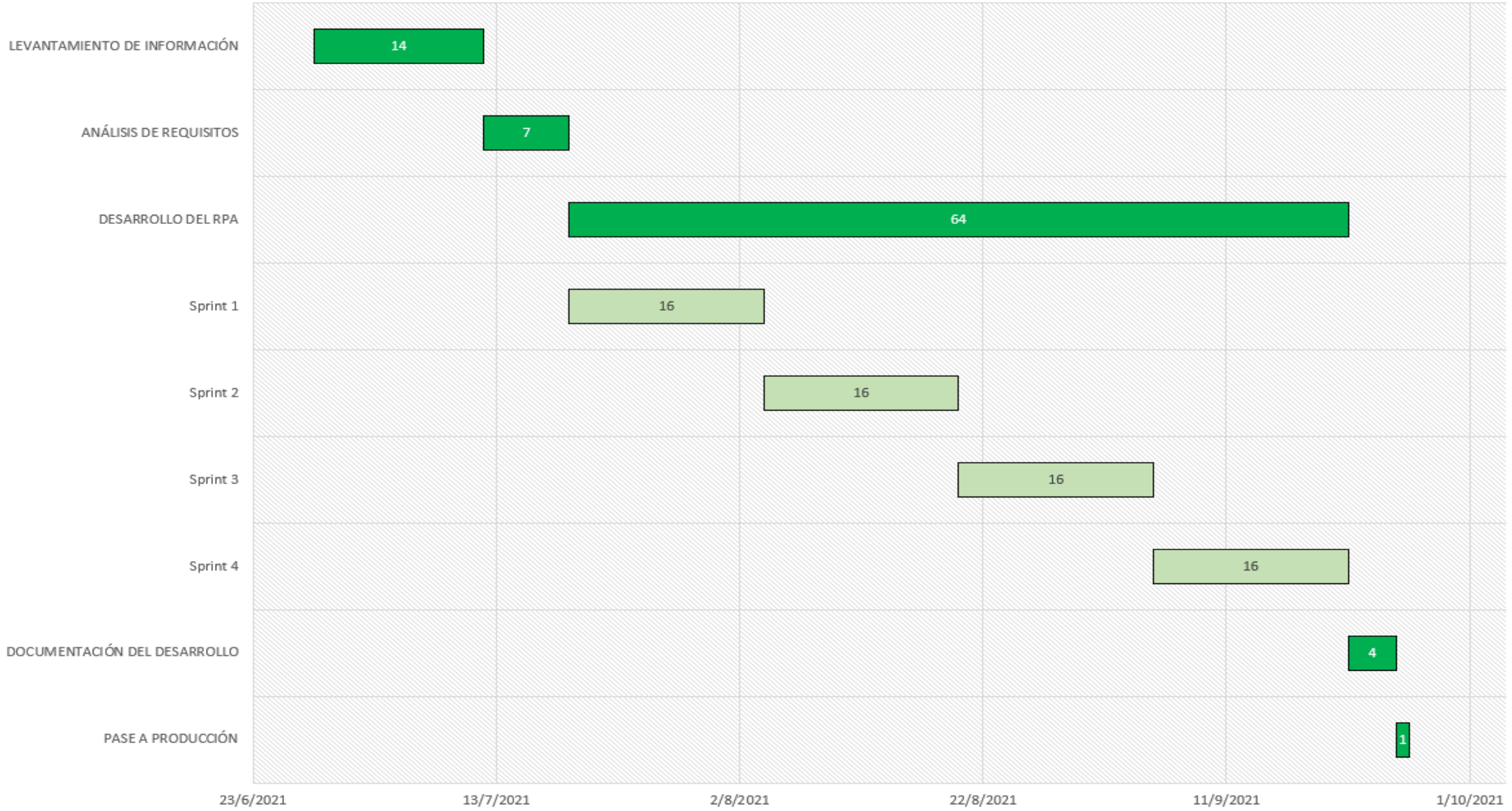
Fuente: Elaboración propia de los autore

Tabla 21: Gantt del desarrollo del Sistema RPA.

Tareas del proyecto	inicio	Fin	Duración (días)
LEVANTAMIENTO DE INFORMACIÓN	2021-06-28	2021-07-11	14
ANÁLISIS DE REQUISITOS	2021-07-12	2021-07-18	7
DESARROLLO DEL RPA	2021-07-19	2021-09-20	64
Sprint 1	2021-07-19	2021-08-03	16
Desarrollo del RPA	2021-07-19	2021-08-01	14
Pruebas de QA	2021-08-02	2021-08-02	1
Pruebas de aceptación	2021-08-03	2021-08-03	1
Sprint 2	2021-08-04	2021-08-19	16
Desarrollo del RPA	2021-08-04	2021-08-17	14
Pruebas de QA	2021-08-18	2021-08-18	1
Pruebas de aceptación	2021-08-19	2021-08-19	1
Sprint 3	2021-08-20	2021-09-04	16
Desarrollo del RPA	2021-08-20	2021-09-02	14
Pruebas de QA	2021-09-03	2021-09-03	1
Pruebas de aceptación	2021-09-04	2021-09-04	1
Sprint 4	2021-09-05	2021-09-20	16
Desarrollo del RPA	2021-09-05	2021-09-18	14
Pruebas de QA	2021-09-19	2021-09-19	1
Pruebas de aceptación	2021-09-20	2021-09-20	1
DOCUMENTACIÓN DEL DESARROLLO	2021-09-21	2021-09-24	4
PASE A PRODUCCIÓN	2021-09-25	2021-09-25	1

Fuente: Elaboración propia de los autores.

Figura 17: Gráfica Gantt del desarrollo del Sistema RPA.



Fuente: Elaboración propia de los autores

Tabla 22: Establecimiento del proyecto.

Sistema de desarrollo	Descripción
Python	Lenguaje de programación Python Versión 3.8.5 Librería, Selenium 3.141.0 IDE: Spyder (Anaconda 3)
SQL Server	Versión 2014 Express

otros	Descripción
Laptop	Procesador: AMD Ryzen 5 3500U RAM instalada: 8.00 GB Sistema operativo de 64 bits, procesador x64 256 SSD

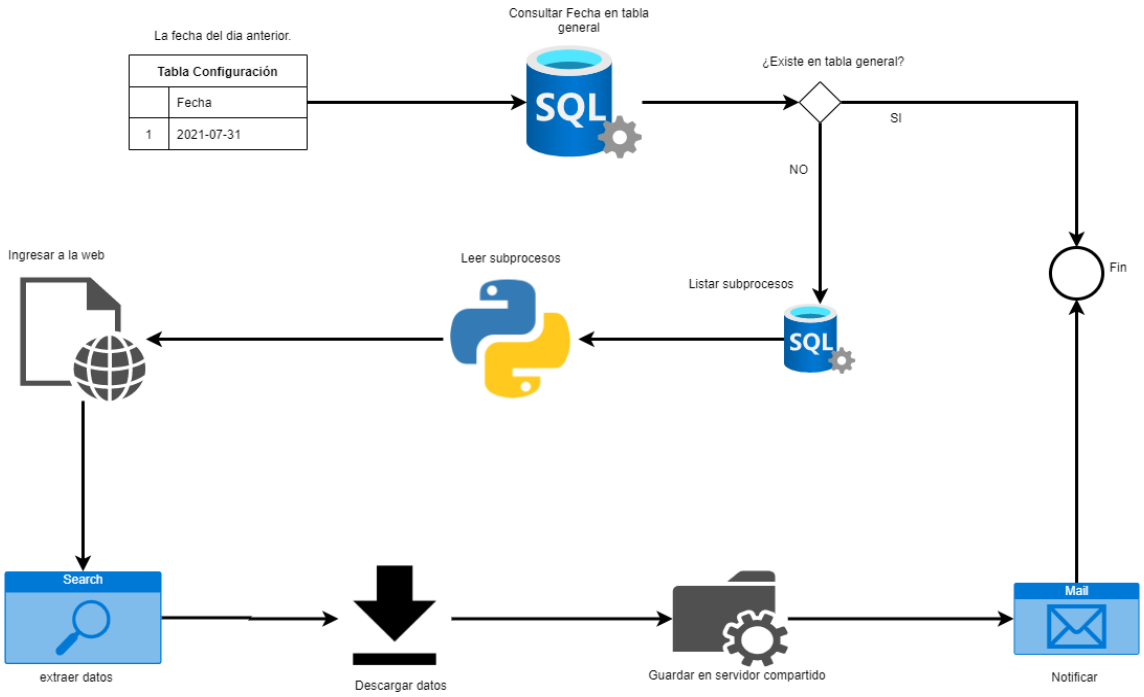
Anexo 15: Arquitectura del RPA.

Consta de una tabla maestra en SQL Server el cual tendrá la configuración de los subprocesos (Nombre subproceso, nombre base de datos, formato de base de datos, Fecha) (ver anexo 10), el campo llamado Fecha que cambiará de acuerdo con el transcurrir de los días (GETDATE () - 1). Cada día se debe descargar datos del día anterior. Además, existe una tabla general que es donde se almacenan día a día los datos extraídos de la web.

Se realizará un cruce de la tabla maestra con la tabla general para saber si la fecha de cada subproceso de la tabla maestra existe en la tabla general.

Esta tabla maestra se encargará de validar si la fecha almacenada de cada subproceso ya existe en la tabla general. En caso existe finaliza el subproceso para evitar duplicidad, de lo contrario se ejecuta el RPA, ingresando a la web para recolectar los datos, almacenarlos y enviar un correo de confirmación de dicha tarea.

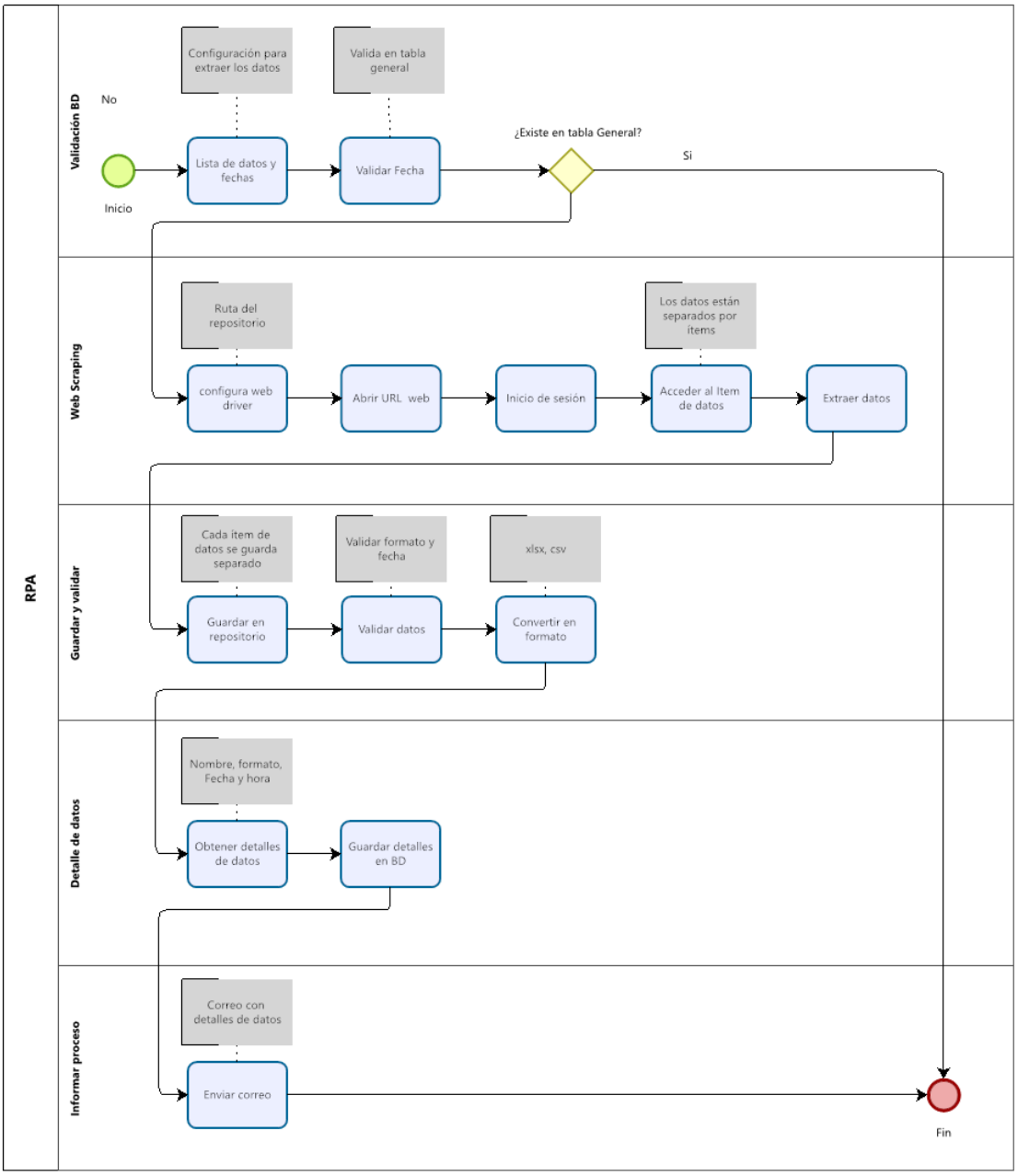
Figura 18: Arquitectura del RPA.



Fuente: Elaboración propia de los autores

El sistema RPA está diseñado para ser ejecutado únicamente en el área de trabajo para el que se tiene destinado, debido a que presenta configuraciones específicas para realizar, además la red tiene restricciones de privacidad.

Figura 19: Flujograma del RPA utilizando técnicas de web scraping.



Fuente: Elaboración propia de los autores

Dentro del proceso RPA con técnicas de web scraping vamos a tener 5 subproceso participantes en el flujo.

1.- **Validación BD:** Este subproceso toma la fecha del día anterior para cada Ítem de datos y valida si ya fue cargado en la tabla general. Si en caso existe ya no sería necesario volver a extraer los datos por que se generaría duplicidad. Una vez que se tenga listo los ítems que están pendiente de extracción pasa al siguiente paso.

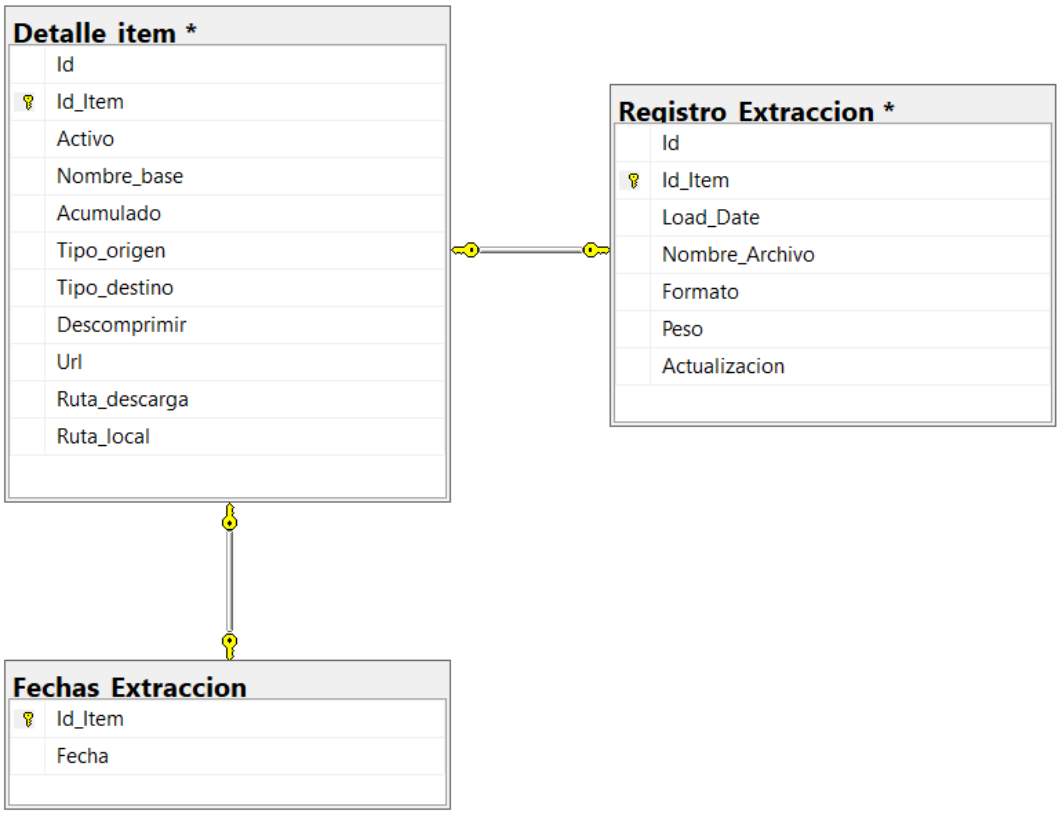
2.- **Web Scraping:** Encargado de realizar el proceso tal cual lo realiza una persona, tales como realizar la conexión y configuración del web driver, asimismo de iniciar sesión y navegar en las diversas pestañas de la web para poder extraer cada uno de los ítems de datos.

3.- **Guardar y validar:** Este subproceso consta de 3 tareas. Lo primero es dirigir cada ítem de datos al repositorio que le corresponde, posteriormente validar si los datos que se han extraído cumplen están de acuerdo con el formato acordado y finalmente si no cumple se encarga de convertirlo el formato correcto.

4.- **Detalle de datos:** Cuenta con funciones capaces de obtener los detalles de cada ítem de datos que fueron extraídos, dentro de los detalles los principales son formato, fecha y hora en el que fue extraído cada ítem.

5: **Informar Proceso:** Este subproceso se encarga de enviar cualquier excepción que tenga el sistema, pero el objetivo principal es enviar un correo llevando una tabla como contenido, en la cual tendremos los detalles de los datos que fueron extraídos, la finalidad de la tabla es brindar mayor detalle y tomar acciones en caso sea necesario.

Figura 20: Estructura de las tablas utilizadas para el proceso de extracción de datos.



Fuente: Elaboración propia de los autores

La tabla **Fechas_Extraccion** guarda el ítem de datos y la fecha correspondientes para poder extraer los datos. El campo fecha es actualizada de manera automática cada día, debido a que en la consulta SQL será de la siguiente manera.

```
SELECT
        Id_Item,
        Fecha = CAST (GETDATE () AS DATE)
FROM Fechas_Extraccion
```

Fechas_Extraccion

Columnas	Descripción
Id_Item	Id del ítem de datos.
Fecha	Fecha para extraer los datos.

En la tabla **Detalle_Item**, encontraremos todo lo necesario para poder extraer los datos de la web, esta tabla guarda la configuración para los ítems, del cual, el RPA tomará como parámetros para extraer y guardar de manera correcta los datos ya definidos tanto en el formato, nombre y ruta donde se debe guardar cada ítem.

Detalle_item

Columnas	Descripción
Id	Id del conjunto de datos
Id_Item	Id del ítem de datos.
Activo	Indica si el ítem debe ser activo o baja.
Nombre_base	Nombre de los datos a extraer en la web.
Acumulado	Indica si es acumulado mensual, quincenal o diario.
Tipo_origen	Tipo de formato de datos según origen.
Tipo_destino	Tipo de formato de datos definidos.
Descomprimir	Se evalúa si el archivo debe ser comprimido o no, según el peso.
Url	Url para acceder a los datos según cada ítem.
Ruta_descarga	Carpeta donde debe ser guardada el archivo de datos, se asigna al web diver.
Ruta_local	Carpeta donde debe ser guardada el archivo de datos

La tabla **Registro_Extraccion**, se encarga de almacenar los detalles o propiedades de cada ítem de datos que se ha extraído de la web. Esta información posteriormente no servirá para validar cuantos datos fueron extraídos en el tiempo oportuno y en el formato correcto. Para realizar esta validación se utilizará la configuración almacenada en la tabla **Detalle_item** utilizando el campo identificador **Id_Item.**

Registro_Extraccion

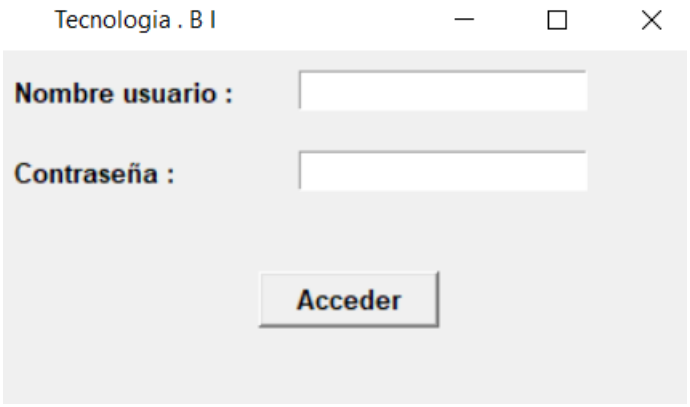
Columnas	Descripción
Id	Id del conjunto de datos
Id_Item	Id del ítem de datos.
Load_Date	Fecha y hora que se registra.
Nombre_Archivo	Nombre con el que se guarda el ítem de datos.
Formato	Formato con el que se guarda el ítem de datos.
Peso	Pero del ítem de datos.
Actualizacion	Fecha y hora de extracción del ítem de datos.

El RPA tiene una tabla de configuración para los subprocesos de los 8 ítems que conforman cada conjunto de datos. Cada subproceso se encarga de recolectar datos distintos y poder exportarlos a un formato estructurado que le corresponde (CSV, XLSX, TXT) (Ver Tabla 13)

Anexo 16: Prototipos de programación del sistema.

Para poder ejecutar de manera manual, el sistema cuenta con un usuario y contraseña de acceso, el cual es administrador por el responsable del área.

Figura 21: Login de ingreso al Sistema RPA.



Fuente: Elaboración propia de los autores

Una vez dentro de la aplicación tendremos opciones de ejecutar que se encargar de realizar la extracción de los datos, enviar reporte se encarga de confirmar mediante correo los datos extraídos, el formato, fecha y hora. El botón salir es para cerrar el sistema.

Figura 22: Menú de opciones del Sistema RPA.

Fuente: Elaboración propia de los autores

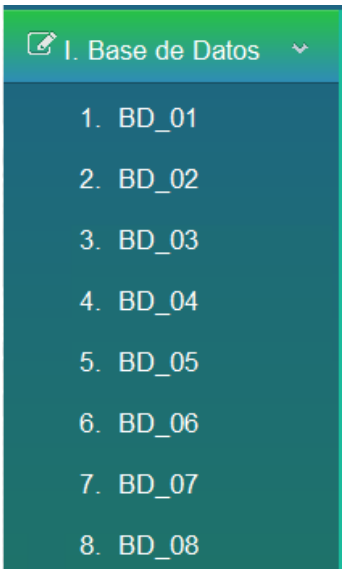
Cuando ejecutamos el RPA, este levanta la web donde por seguridad también tiene usuario y contraseña, en esta web el RPA de manera automática completa los datos configurados y da clic en el botón ingresar.

Figura 23: Login de ingreso a la web para la extracción de datos.

Fuente: Elaboración propia de los autores

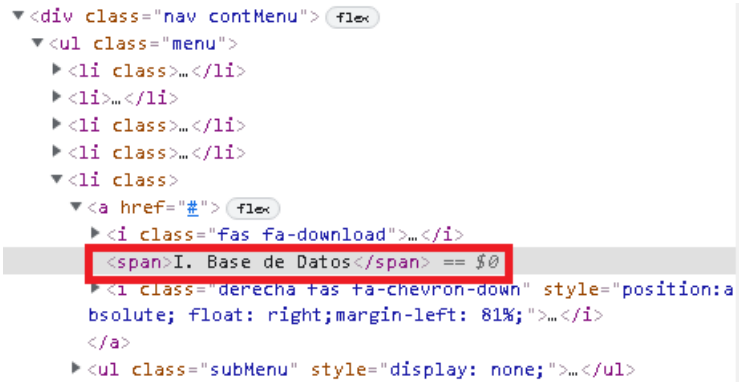
Acceder al apartado de descargas de los datos. En este apartado tenemos los datos que necesitamos recolectar dispersos en diferentes ítems. El RPA debe iterar con la tabla de configuración antes mencionada.

Figura 24: Apartado de los ítems de datos



Fuente: Elaboración propia de los autores

Para realizar esto, el RPA debe interactuar con las clases y demás elementos dentro del HTML de la página.



Realizar la extracción de los datos, que posteriormente será llevado a un dataframe de Python y pasado a una base datos del tipo definido por el usuario.

Figura 25: Tabla de datos alojado en la web

session_	fecha_carga	creation_date	close_date	case_of_use_	resol	priorit
00VNHVR2OGDTF3MT3(	3/11/2021	1/11/2021 18:29	1/11/2021 18:34	Inactiv	Si	3
01TH6JZFRNR2UZQAQQI	3/11/2021	1/11/2021 16:43	1/11/2021 16:45	Inactiv	Si	41
0304FVY2RKRK3EHQWEI	3/11/2021	1/11/2021 15:40			Si	365
03I1LQFXIKY0PRGLCDMI	3/11/2021	1/11/2021 18:49	1/11/2021 18:57	Coach	No	4
061OAOLJPY8JFVAQEGW	3/11/2021	1/11/2021 22:06	1/11/2021 22:25	Coach	Si	145
06OPVA8V8YY6BJPR731(	3/11/2021	1/11/2021 15:47				385
073WEI82NAQDGXIV08J	3/11/2021	1/11/2021 21:03			No	4
0DJOA6XUMCBPBCAEW	3/11/2021	1/11/2021 12:49			Si	465
0GQR467FGCSG4QW6W	3/11/2021	1/11/2021 12:42	1/11/2021 12:54	Inactiv	Si	371
0O2MQZL7EG4ER2EII85	3/11/2021	1/11/2021 14:50	1/11/2021 15:08	Inactiv	No	
0PB8PVBOENXAT2WE6C	3/11/2021	1/11/2021 19:12	1/11/2021 22:09		Si	425
0PB8PVBOENXAT2WE6C	3/11/2021	1/11/2021 22:07	1/11/2021 22:09	Inactiv	Si	425
0QVCVI0ZES3X7NEPQGV	3/11/2021	1/11/2021 18:41	1/11/2021 18:44	Inactiv	No	152
0QXLD4TP6ZKZMTAIZFB	3/11/2021	1/11/2021 19:42			No	3

El RPA esa configurado de acuerdo con la estructura del HTML de la tabla alojada en la web, en cada línea de la etiqueta "**tr**" contiene los datos que necesitamos capturar, luego podemos ubicar la etiqueta a través del método de posicionamiento de selenium y obtener los datos

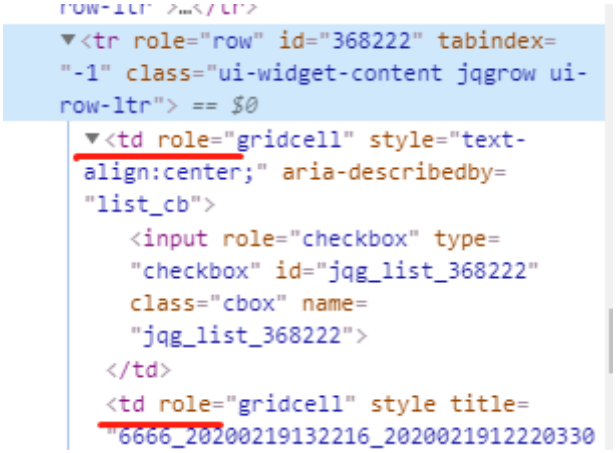


Figura 26: Extracción de datos de las etiquetas

```
def find_all_data(driver,page):
    dat = []
    lst = []
    for i in range(1, page + 1):

        inputpage = driver.find_element_by_xpath("//*[@id='pager_center']/table/tbody/tr/td[2]/input")

        inputpage.clear()

        inputpage.send_keys(str(i))

        inputpage.send_keys(Keys.ENTER)

        time.sleep(3)
        element = driver.find_element_by_css_selector('#list')
        tr_contents = element.find_elements_by_tag_name('tr')
        for tr in tr_contents:
            for td in tr.find_elements_by_tag_name('td'):
                lst.append(td.text)
            dat.append(lst)
            lst = []
    return dat
```

Una vez recuperados los datos, es hora de insertarlos a un dataframe y posteriormente esto será depositado en la carpeta que le corresponde con el formato de datos especificados.

Figura 27: Almacenar datos en el formato definido

```
lst = find_all_data(driver,3)
time.sleep(2)
driver.close()

workbook = xlsxwriter.Workbook('D:Item_datos_01.xlsx')

worksheet = workbook.add_worksheet()

coloum = len(lst)
row = len(lst[0])
print(coloum,row)
for i in range(0,coloum):
    for j in range(0,row):
        text = lst[i][j]

        worksheet.write(i,j,text)
workbook.close()
```

Una vez se finalice el proceso, el RPA enviará un correo detallando los datos extraídos.

Figura 28: Notificación por correo de los datos extraidos.

Se encuestran disponible las siguientes bases.

Item	NOMBRE_ARCHIVO	Actualizacion
1	BD_01.xlsx	2021-10-30 09:01
2	BD_02.xlsx	2021-10-30 09:03
3	BD_03.xlsx	2021-10-30 09:10

Saludos.

Fuente: Elaboración propia de los autores

Módulo de alerta de errores mediante correos, se encarga de alertar cada una de las excepciones que tenga el sistema, entre los más comunes son problemas para conectarse a la base de datos y problemas de conexión con el driver del navegador. También se encarga de confirmar que se recolectó de manera correcta los datos de cada subproceso.

Configuración de conexión al servidor, se utiliza para consultar la tabla maestra y la tabla general previo a la recolección de datos.

A continuación, se detalla el cronograma en el cual se realizará la extracción de los datos.

Figura 29: Cronograma de ejecución del RPA.

Frecuencia	Hora	Estado
Diaria	07:30 a. m.	Habilitado
Diaria	08:30 a. m.	Habilitado
Diaria	09:30 a. m.	Habilitado
Diaria	10:30 a. m.	Habilitado
Diaria	02:30 p. m.	Habilitado

Fuente: Elaboración propia de los autores