Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Profesional de Estadística

Análisis de Series de Tiempo y Machine Learning para proyectar una eficiente gestión de subsidios ante EsSalud

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Licenciada en Estadística

AUTOR

Keyla Fiorela VALVERDE SHUAN

ASESOR

Mg. Emerson Damian NORABUENA FIGUEROA

Lima, Perú

2023

Referencia bibliográfica

Valverde, K. (2023). *Análisis de Series de Tiempo y Machine Learning para proyectar una eficiente gestión de subsidios ante EsSalud*. [Trabajo de Suficiencia Profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Keyla Fiorela Valverde Shuan
Tipo de documento de identidad	DNI
Número de documento de identidad	72879069
URL de ORCID	https://orcid.org/0009-0002-5174-6163
Datos de asesor	
Nombres y apellidos	Emerson Damian Norabuena Figueroa
Tipo de documento de identidad	DNI
Número de documento de identidad	45259683
URL de ORCID	https://orcid.org/0000-0003-2909-7080
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Zoraida Judith Huamán Gutiérrez
Tipo de documento	DNI
Número de documento de identidad	09890094
Miembro del jurado 1	
Nombres y apellidos	Hugo Marino Rodríguez Orellana
Tipo de documento	DNI
Número de documento de identidad	40162362
Datos de investigación	
Línea de investigación	A.3.2.6. Análisis de Datos y Modelamiento de Problemas de la Sociedad

Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	Edificio: Universidad Nacional Mayor de San Marcos País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Latitud: -12.0561 Longitud: -77.0845
Año o rango de años en que se realizó la investigación	Enero 2023 - Mayo 2023
URL de disciplinas OCDE	Estadísticas, Probabilidad https://purl.org/pe-repo/ocde/ford#1.01.03 Ciencias de la información https://purl.org/pe-repo/ocde/ford#1.02.02

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú. Decana de América
FACULTAD DE CIENCIAS MATEMÁTICAS
ESCUELA PROFESIONAL DE ESTADISTICA

ACTA DE SUSTENTACIÓN DEL TRABAJO DE SUFICIENCIA PROFESIONAL PARA LA OBTENCIÓN DEL TÍTULO PROFESIONAL DE LICENCIADA EN ESTADÍSTICA
(PROGRAMA DE TITULACIÓN PROFESIONAL 2023)

En la UNMSM – Ciudad Universitaria – Facultad de Ciencias Matemáticas, siendo las 9:30 horas del sábado 21 de octubre del 2023, se reunieron los docentes designados como Miembros del Jurado Evaluador (PROGRAMA DE TITULACIÓN PROFESIONAL 2023): Dra. Zoraida Judith Huamán Gutiérrez (PRESIDENTE), Mg. Hugo Marino Rodríguez Orellana (MIEMBRO) y el Mg. Emerson Damián Norabuena Figueroa (MIEMBRO ASESOR), para la sustentación del Trabajo de Suficiencia Profesional titulado: "**ANÁLISIS DE SERIES DE TIEMPO Y MACHINE LEARNING PARA PROYECTAR UNA EFICIENTE GESTIÓN DE SUBSIDIOS ANTE ESSALUD**", presentado por la señorita **Bachiller KEYLA FIORELA VALVERDE SHUAN**, para optar el Título Profesional de Licenciada en Estadística.

Luego de la exposición del Trabajo de Suficiencia Profesional, la Presidente invitó a la expositora a dar respuesta a las preguntas formuladas.

Realizada la evaluación correspondiente por los Miembros del Jurado Evaluador, la expositora mereció la aprobación .....Bueno....., con un calificativo promedio de Dieciseis (16)

A continuación, los Miembros del Jurado Evaluador dan manifiesto que la participante **Bachiller KEYLA FIORELA VALVERDE SHUAN**, en vista de haber aprobado la sustentación de su Trabajo de Suficiencia Profesional, será propuesta para que se le otorgue el Título Profesional de Licenciada en Estadística.

Siendo las 10:00 horas se levantó la sesión firmando para constancia la presente Acta.

Dra. Zoraida Judith Huamán Gutiérrez
PRESIDENTE

Mg. Hugo Marino Rodríguez Orellana
MIEMBRO

Mg. Emerson Damián Norabuena Figueroa
MIEMBRO ASESOR

CERTIFICADO DE SIMILITUD

Yo, Emerson Damián Norabuena Figueroa en mi condición de asesor acreditado con Resolución Decanal N° 001617-2023-D-FCM/UNMSM del Trabajo de Suficiencia Profesional, cuyo título es "ANÁLISIS DE SERIES DE TIEMPO Y MACHINE LEARNING PARA PROYECTAR UNA EFICIENTE GESTIÓN DE SUBSIDIOS ANTE ESSALUD", presentado por la bachiller KEYLA FIORELA VALVERDE SHUAN, para optar el título de Licenciada en Estadística.

Certifico que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante el software de similitud textual, el documento evaluado cuenta con el porcentaje de **6%** de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional**.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del título correspondiente.

_____

DNI: 45259683

Mg. Emerson Damián Norabuena Figueroa

Huella Digital

Resumen

El presente estudio tiene como finalidad reconocer los factores o variables que generan una buena gestión de subsidios antes Essalud, gestionado por la empresa HumaSer. Para ejecutar dicho objetivo contamos con el apoyo de un sistema que nos brinda una supervisión interna y diligencias de gestión, con la finalidad de salvaguardar los recursos contra pérdidas por ineficacias operativas mediante Essalud; comenzamos a reconocer los primordiales riesgos en el largo proceso y luego mejoramos la gestión; ejecutando acciones que nos ayude a disminuir y prevenir la posibilidad de los montos no recuperados en cada una de las cuentas adquiridas por los clientes, previniendo retrasos, pérdidas, y obedeciendo las normas interpuestas.

Esta investigación está elaborada con los registros obtenidos por cada uno de los clientes, dicha información fue analizada a fin de conocer el estado en el cual se encuentran los tramites de subsidios por incapacidad y bajo esa perspectiva desarrollar un plan de trabajo.

Ante ello se realiza el uso de la técnica de Regresión lineal múltiple (Machine Learning – supervisado) y el modelo de series de tiempo, propuestos por Box – Jenkins con el propósito de encontrar variables que nos indique o nos brinde una administración eficiente en la gestión de subsidios y por otro lado, visualizar el comportamiento de la data a través del tiempo con la finalidad de realizar pronósticos adecuados. Esta técnica nos ayudará alcanzar nuestro objetivo, debido a que son utilizadas para identificar outliers, correlaciones y explicar la influencia que genera las variables predictoras en nuestra variable dependiente en una gran cantidad de datos y que nos permita predecir los resultados para llegar a mejorar la efectividad en la toma de decisiones

Palabras clave: Regresión, Incapacidad temporal, anomalías, series de tiempo, ARIMA.

Abstract

The purpose of this study is to recognize the factors or variables that generate a good management of subsidies before Essalud, managed by the company HumaSer. To execute this objective we have the support of a system that provides us with internal supervision and management procedures, in order to safeguard resources against losses due to operational inefficiencies through Essalud; we start to recognize the main risks in the long process and then we improve the management; executing actions that help us to reduce and prevent the possibility of unrecovered amounts in each of the accounts acquired by clients, preventing delays, losses, and obeying the interposed regulations.

This investigation is elaborated with the records obtained by each one of the clients, said information was analyzed in order to know the state in which the procedures for disability subsidies are found and under this perspective develop a work plan.

Given this, the use of the Multiple Linear Regression technique (Machine Learning - supervised) and the time series model, proposed by Box - Jenkins, are made with the purpose of finding variables that indicate or provide us with an efficient administration in the management of subsidies and on the other hand, to visualize the behavior of the data over time in order to make adequate forecasts. This technique will help us achieve our objective, because they are used to identify outliers, correlations and explain the influence generated by the predictor variables in our dependent variable in a large amount of data and that allows us to predict the results to improve effectiveness. in making decisions

Keywords: Regression, temporary incapacity, anomalies, time series, ARIMA.

Contenido

Índice de tablas

Índice de Figuras

I. Introducción

En la Conferencia Internacional del Trabajo N° 89, se abarcó el tema de la seguridad social y su contenido, se afirmó que " los sistemas de seguridad social deben incluir al menos las siguientes contribuciones de salud, pensiones y asistencia: Atención en medicina integrativa; indemnización por afección o alguna contingencia debidamente certificada; contribuciones de salud y económicos de la maternidad; beneficios económicos de salud a largo plazo derivados de la vejez; invalidez y/o pervivencia; prestaciones económicas por enfermedad profesional, accidente de trabajo y condición de desempleo (Organización de las Naciones Unidas [ONU],1948).

Después en 1948, la asamblea General de las Naciones Unidas admitió la declaración Universal de los Derechos Humanos. El artículo 22 del citado documento ordena que todo individuo posee derechos a la seguridad social y por medio de los esfuerzos nacionales y la asistencia internacional en materia de organización y recursos de la previsión social para lograr la satisfacción del derecho social, cultural y económico (Campos ,2010).

El Seguro Social de Salud (ESSALUD,2012) sostiene como responsabilidad brindar garantía a todos sus asegurados a través de las prestaciones de salud, social, beneficios y económicas. En este último caso, Essalud asume la responsabilidad del asegurado después de los 20 primeros días de descanso médico brindados por el médico a cargo.

En la siguiente investigación se tiene como objetivo reconocer los factores o variables que nos brinde un adecuado manejo de subsidios ante Essalud, dicha labor se efectúa en una empresa REMYPE que ofrece asistencia en soluciones de gestión humana como recuperos, reembolsos, control, validación y canjes de descansos médicos por incapacidad temporal. La

empresa inició su funcionamiento hace 20 años, es independiente en su organización y gobierno de datos, la cual consiste en trabajar con los registros que brinda cada institución; cuenta con sistema implementado para tener un buen desarrollo en la base de datos hacia los reportes y análisis que se realizará en el área de Data.

Dicha empresa cuenta con distintas áreas, pero, el área de Data es la encargada de ofrecer soporte informativo de los avances a los lideres implicados en la toma de decisiones. Habiendo colaborado en la mencionada área, cuyas funciones son control, administración, procesamiento y elaboración de reportes estadísticos de las bases de datos.

La empresa tiene como finalidad brindar a sus clientes un resultado efectivo en el manejo adecuado de los subsidios y realizar una adecuada gestión; por ello, con el objetivo de proporcionar una mayor efectividad de montos reconocidos ante Essalud para sus clientes, se desea detectar las variables que aumenten la efectividad del manejo adecuado de los subsidios. Mediante la técnica de regresión lineal múltiple y el modelo ARIMA se pretende encontrar las variables que nos llevará a realizar un adecuado manejo de gestión en subsidios del que ya se tiene, para así poder seguir brindando alta efectividad en los resultados y por otra parte realizar un pronóstico para tener un futuro panorama y establecer decisiones adecuadas.

II. Descripción de la Actividad

La empresa encargada de realizar gestiones humanas cuenta con experiencia y acreditaciones en corporaciones nacionales y multinacionales. Ante la alta demanda de solicitudes de reembolso, por parte de los clientes. La empresa implementa como nuevo proyecto, generar estrategias para ingresar las solicitudes de reembolsos correspondientes y a su vez tener éxito en cada una de ellas.

Por tal razón se implementa una nueva estrategia, haciendo uso de la minería de datos para ubicar las variables que nos ayuden a realizar una gestión de subsidio eficiente y de calidad.

2.1. Datos de la empresa o institución

2.1.1. Nombre de la institución

HumaSer

2.1.2. Periodo de duración del TSP

Del 01 de Diciembre del 2022 el 31 de Mayo 2023

2.1.3. Razón Social

Humana Resultados Eficaces S.A.C

2.1.4. Dirección

Jirón José Cossio Nro. 228-Interior 201 Magdalena.

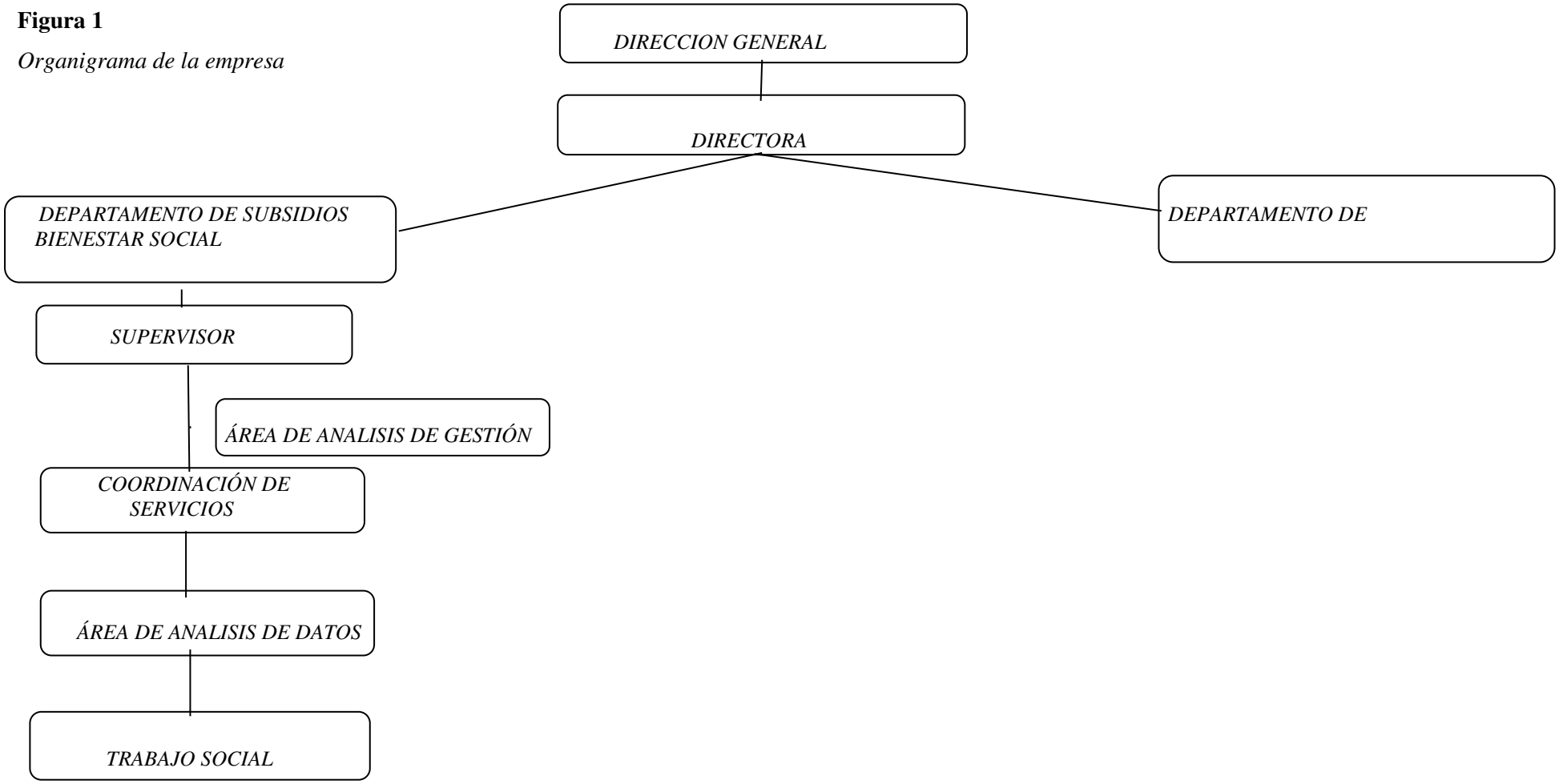
2.1.5. Correo electrónico

Keyla.valverde@humaser.net

2.1.6. Organigrama de la empresa

Figura 1

Organigrama de la empresa



Nota: HumaSer.

2.2. Finalidad y Objetivos de la empresa

2.2.1. Finalidad

La empresa encargada de realizar gestiones humanas tiene como propósito brindar soluciones de gestión humana, como son los subsidios o reembolsos ante Essalud; ofreciendo de manera efectiva resultados en los clientes.

Objetivo: Atender a los clientes de manera efectiva y con los mejores resultados en cada uno de sus requerimientos.

Misión: Brindar soporte a cada uno de nuestros clientes durante el proceso de subsidios antes Essalud.

Visión: Convertirnos en la empresa de gestión humana, con el más alto nivel de calidad y atención personalizada.

2.3. Descripción de la actividad

La entidad privada cuenta con un área encargada en realizar el análisis de datos, cuyo propósito es:

Determinación de estrategias para cumplir las metas establecidas.

Ayudar a reconocer y solucionar problemas específicos, haciendo uso del análisis de datos.

Realizar análisis predictivos para determinar el futuro de la empresa.

Visualización de los datos para comprender los resultados actuales y tendencias futuras.

Colaborar con las gestoras con la información de diversos clientes dada la necesidad.

Realizar la extracción de los resultados de cada semana y proporcionar las posibles soluciones para una buena gestión.

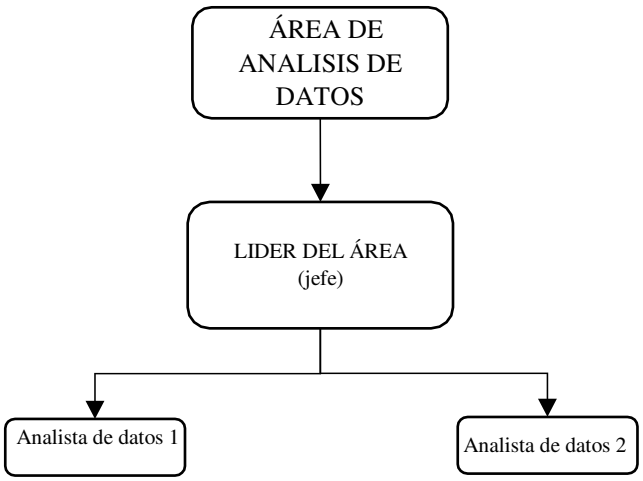
Agrupar, filtrar e interpretar los datos obtenidos semanalmente.

Detectar anomalías en la base de datos y generar reporte de incidencias.

2.3.1. Organigrama del área

Figura 2

Organigrama del área



Nota: Estructura del área de análisis de datos. Fuente: HumaSer.

2.3.2. Finalidad del trabajo

La investigación realizada tiene como designio, realizar un modelo adecuado seleccionando las variables predictoras adecuadas que nos permita identificar los factores que influyen en una buena gestión de subsidios, es decir, los reembolsos que nos brinda Essalud por las gestiones realizadas.

2.3.3. Objetivos del Trabajo

Objetivo general

Identificar las variables que permitan realizar una buena gestión de subsidios ante Essalud.

Objetivos específicos

Pronosticar el monto que se presentará ante Essalud para mayo del 2023.

Hallar el mejor modelo de Regresión lineal múltiple para pronosticar el monto reconocido por Essalud.

Determinar la relación entre el monto presentado y el monto reconocido por Essalud.

2.3.4. Problemática

En la actualidad la empresa posee diversos clientes y se conoce que cada cliente tiene una información y clasificación de manera distinta ante los estatus de reembolso antes Essalud, a pesar de tener una estrategia establecida, no satisface la alta demanda de registros adquiridos por los clientes y por ende el trámite de reembolso no es eficiente. Por tal razón se busca conocer una mejor estrategia que nos permita realizar un adecuado subsidio ante Essalud.

2.3.5. Metodología y procedimientos

Metodología

Tipo de investigación: Concorde con Hernández et al. (2014), esta investigación es de enfoque cuantitativo por lo que se desea calcular y medir el grado de correlación de las variables de estudio, por su clasificación es de tipo básico. Según su alcance es descriptivo, es decir busca describir y visualizar el comportamiento de los datos.

Diseño de investigación: Es no experimental, ya que se utiliza la información existente para realizar dicho estudio y transversal porque la información fue obtenida en un solo tiempo (Hernández et al.,2014).

Población, muestra y unidad de análisis

En el presente estudio se tiene como población a todos los registros obtenidos en el periodo enero del 2019 a mayo del 2023. No se realizó una muestra porque dicho estudio se encuentra enfocado a trabajar con toda la población de interés. Siendo nuestra unidad de análisis cada uno de los registros obtenidos en el periodo enero del 2019 a mayo del 2023.

Instrumentos de recolección de datos

Fuentes:

Los datos o registros fueron usados del Formulario HumaSer cuyo periodo comprende entre enero del 2019 hasta mayo del 2023.

Instrumento:

El principal instrumento, que nos brinda la información de los datos registrados de cada cliente se encuentra en el sistema de la empresa.

Plan de procesamiento y análisis estadísticos de datos

A fin de realizar el desarrollo del análisis de los datos, utilizaremos el lenguaje de programación Python 3 y cuyos procedimientos a seguir para la elaboración de dichos análisis son:

Entendimiento del negocio.

Entendimiento de la data.

Preparación de los datos.

Modelado y validación del modelo.

III. Marco teórico

3.1. Antecedentes

3.1.1. Antecedentes Nacionales

Sánchez (2022) nos menciona en su investigación titulada “Proceso de análisis jerárquico y regresión lineal múltiple para la priorización de estrategias de negocio en una central de riesgo en Lima”. Perú. Objetivo: Emplear un modelo estadístico y a su vez el desarrollo de Análisis Jerárquico. Metodología: Dicha investigación es de tipo cuantitativa y según su nivel de investigación es descriptiva. Instrumento: Se empleó entrevistas a cada uno de los expertos Stakeholders, basándose en un cuestionario que tenga la finalidad de realizar una medición de criterios. Conclusión: Al realizar la aplicación del modelo y al mismo tiempo el desarrollo del estudio jerárquico se concluyó en enfocarse como primer punto la elaboración de estrategias de financiamiento, ya que brindará rentabilidad teniendo en cuenta el presupuesto otorgado por cada productor analítico desarrollado. La estrategia de penetración que tiene como finalidad penetrar mercado y precios; la estrategia de planificación fue la tercera y última en ser priorizada, pues comercializa los productos analíticos para dicho negocio como lo es una Central de Riesgos.

Lizares (2017) cuya tesis titulada “Comparación de Modelos de clasificación: regresión logística y Árboles de clasificación para evaluar el rendimiento académico". Perú. Objetivo: Es realizar la comparación de dos modelos, tales son: regresión logística y árboles de clasificación, cuya finalidad es cualificar el desempeño en la formación académica de manera adecuada. Metodología: el estudio ejecutado es descriptivo y transversal de diseño observacional, no experimental. Instrumento: IBM SPSS statistics. Conclusión: el modelo de regresión logística mostró las variables más influyentes con un p-valor menor al 5%, por otro lado, el modelo de árboles de decisión

indica que las variables predictoras principales son las horas de estudios empleadas, las diferentes instituciones educativas y el placer por llevar los cursos. por ende, se da a conocer como el modelo adecuado.

En la tesis elaborada por Castillo (2022) titulada "Desarrollo de modelos predictivos de regresión en la industria minera mediante el uso de algoritmo de machine learning". Perú. Objetivo: Elaborar modelos de pronóstico de regresión para una industria minera haciendo uso de machine learning. Metodología: La investigación elaborada es no experimental cuantitativo, de tipo exploratorio y de diseño transversal. Instrumento: La información obtenida proviene de la plataforma institucional Conclusión: Se realizaron diferentes modelos de machine learning de las cuales se identificaron los modelos adecuados ante los problemas más robustos, es así que el modelo adecuado para evaluar el precio del oro se optó por la técnica SVR con un R2=0.94, MAE=4.63, RMSE=5.29; el modelo adecuado para valorar el sílice en el concentrado de hierro es Gradient Boosting Regressor con un R2=0.51, MAE=0.81, RMSE=0.81 y como último modelo se tuvo a Random Forest Regressor con un R2=0.98, MAE= 0.87, RMSE=0.91, siendo este último el modelo que ayudará a pronosticar la consumición de los combustibles en los vehículos de carga pesada.

Quispe (2018) en el trabajo de suficiencia profesional titulado "Implementación de un modelo predictivo para incrementar la captación de seguros en una entidad financiera". Perú. Objetivo: Elaborar un modelo predictivo que permita realizar la recaudación de asegurados para Protecciones Múltiples cuya finalidad es ampliar las ventas y obtener la permanencia en los seguros por un tiempo mínimo de 3 meses. Metodología: El tipo de investigación es Explicativa debido a que permite interpretar el significado de cada variable a predecir y predictora. Instrumento: Los registros históricos almacenados en la plataforma de la empresa. Conclusión:

Haciendo uso del modelo de regresión logística se identificaron 9 variables la principal característica del modelo alcanza el 87,7% de rendimiento, lo que lo convierte en un modelo aún asequible. Con base en estos resultados, el modelo se implementó en una campaña comercial en 2016 para vender directamente a los clientes más interesados para lograr sus objetivos: las ventas mensuales de pólizas aumentaron a 710, la efectividad de la campaña superó el 2.7% y la tasa de fuga disminuyó en un 15% % al 8%.

3.1.2. Antecedentes Internacionales

Bastidas y Bernard (2020). En el artículo titulado "Adherencia al tratamiento en diabetes tipo 2: un modelo de regresión logística. Caracas 2017-2018". Venezuela. Objetivo: Determinar cómo las variables de interés ayudan a pronosticar y catalogar a las personas vinculadas o no al tratamiento. Metodología: La investigación es no experimental puesto que se usa la información existente para el desarrollo del estudio, por otro lado, es de diseño transeccional - causal. Instrumento: Aplicaron un cuestionario con tres dimensiones respecto a la depresión, el estilo de vida y el aspecto emocional. Conclusión: Respecto al modelo planteado se dice que fue significativo, es decir se encontró que las mujeres con mayor tiempo en tener diabetes tienen mayor probabilidad de evidenciarse adherencias al tratamiento. Se concluye la importancia de los aspectos sociodemográfico, características relacionadas con la enfermedad y aspectos psicológicos para ayudar a diseñar intervenciones en la región y fortalecer un encuadre multidisciplinario para el desarrollo del tratamiento de dicha enfermedad.

López et al. (2018) cuyo artículo titulado "Análisis de los hurtos en Colombia durante el año 2017 mediante los modelos de regresión lineal múltiple y la regresión ponderada geográficamente". Colombia. Objetivo: Estudiar la relación entre los componentes socioeconómicos y los robos en distintas ciudades en el país de

Colombia para el año 2017. Metodología: El estudio realizado es de tipo cuantitativo - correlacional. Instrumento: La base de datos obtenidas proceden de las diversas entidades gubernamentales como institución cohesionada de la Policía Nacional de Colombia. Conclusión: Se probaron variables como matrículas en universidades por mil personas, evaluación por régimen de participación y población adscrita en el estrato urbano para explicar el 69,5% de variación. Los registros de robo personal y de teléfonos celulares se estimaron a nivel mundial haciendo uso del modelo de regresión lineal múltiple para 532 ciudades y el 50,16 % para el modelo estadístico de regresión ponderada geográficamente que ignora las categorías de ciudades. Este modelo tiene los coeficientes ligeramente diferentes respecto al nivel de ciudades, lo que refleja el efecto de heterogeneidad económica y también el aspecto social en las tasas de robo a nivel nacional.

Sommerfeld (2020) en su investigación "Inteligencia emocional y estrés laboral en docentes de educación escolar básica durante la pandemia covid-19". Paraguay. Objetivo: plantea establecer niveles de inteligencia emocional, para identificar el nivel de estrés que se existe en la labor de los maestros de educación primaria del primer y segundo ciclo de Capitán Meza – Itapuá – Paraguay. Metodología: En un estudio descriptivo, no experimental - transversal. Instrumento: emplearon una triangulación concurrente (DITRIAC) para obtener y analizar los datos de una muestra de 12 docentes, 6 mujeres y 6 varones .Conclusión: Obtuvo como resultado que el grado de la inteligencia emocional predomina en maestros de primer y segundo ciclo ya que 10 de los 12 sujetos obtuvieron un alto y medio de puntuación en su Cociente General (CG) y también está presente una correlación positiva (r=0.67) en la inteligencia emocional entre el estrés laboral es decir a mayor puntuación en CG se presenta menor estrés.

Cruz et al. (2023) su estudio titulado “Estimación de las tendencias de precios del agave mezcalero en México utilizando modelos de regresión lineal múltiple”. México. Objetivo: desarrollar un modelo para estimar los precios medios rurales (PRM) en México con información tomada del periodo 1999-2018. Metodología: cuyo estudio es descriptivo y no experimental. Instrumento: se obtuvo la información del sistema encargado de realizar el almacenamiento de los precios del agave mezcalero. Conclusión: Se identificaron las variables con influencia significativa en la determinación de la ARP: el rendimiento de Agave Mezcalero, la ARP de Agave Tequilero y la nueva superficie sembrada de Agave Tequilero con un ajuste de 6 periodos. En general, se generaron tres modelos: El modelo 2 se consideró el más adecuado porque permite realizar pronósticos a futuro con la nueva superficie sembrada de Agave Tequilero con 2 variables independientes. YAM y NPAATt-6 fueron útiles para predecir el 65,5% de las variaciones anuales de la ARP y ayudaron a reconocer la tendencia negativa del precio del Agave de 2020 a 2024. Por lo tanto, el uso del MLRM para estimar la ARP del agave puede ser una herramienta importante para el desarrollo de la predicción en el comportamiento del cultivo.

Ortiz Cardona, (2020) efectuó un estudio analítico cuyo objetivo es establecer el comportamiento futuro de la propagación del COVID-19, dan a conocer que la pandemia ha generado una crisis significativa en los hospitales, colegios y el sector alimentario. Esta información fue tomada de todos los casos positivos registrados en Colombia desde el 6 marzo hasta el 28 de octubre. Como resultado se obtuvo un p-valor de 0.01, es decir la serie de tiempo cumplió con el comportamiento estacional y el modelo adecuado para este pronóstico fue ARIMA (11, 1,13) debido a que el pronóstico elaborado tiene una aproximación adecuada al comportamiento de la serie de tiempo real, se concluyó que debido a la gran variabilidad y fluctuación que

presento la serie de estudio dificultaba que el modelo ARIMA se adecue al comportamiento de esta serie de tiempo.

3.2. Bases teóricas

3.2.1. Series de Tiempo

ofrecen criterios para abordar problemas de estimación basados en registros de tamaño limitado. El término "series de tiempo" se refiere a una recopilación de datos de una variable a lo largo de un intervalo de tiempo definido, manteniendo un orden o secuencia específica, como días, meses o años. Los datos recopilados siempre se expresan como números cambiantes a lo largo del tiempo. Al llevar a cabo un análisis, es fundamental examinar inicialmente su tendencia y estacionalidad. Esto sienta las bases para emplear técnicas analíticas que permitan una exploración exhaustiva. El objetivo final es generar proyecciones para futuros acontecimientos (Garcia,2010).

3.2.2. Clasificación de las series de tiempo

Las series de tiempo se clasifican en estacionarias y no estacionarias*:*

Estacionarias: se refiere a una serie que muestra constancia a lo largo del tiempo, lo que implica que sus propiedades como la media, varianza y covarianza permanecen constantes en todo el período. Se observa que los valores dentro de la serie tienden a fluctuar alrededor de su valor medio (Gonzales, 2009).

No estacionarias: Dentro de esta clasificación, las propiedades estadísticas como la media, la varianza y la covarianza no permanecen constantes. En consecuencia, en una serie de tiempo no estacionaria, la media no muestra una constancia; más bien, define una dirección de crecimiento o decrecimiento a lo largo de un período extenso. Esto resulta en que la serie no se mantiene en torno a un valor invariable,

sino que revela una tendencia discernible a lo largo del tiempo (Banco central de Reserva del Perú, 2017).

3.2.3. Metodología Box-Jenkins

El método ARIMA fue propuesto en 1976 por Box y Jenkins, por lo que lleva sus nombres. Para aplicar este enfoque, es necesario contar con una serie temporal que incluya datos numéricos en diferentes intervalos, como diarios, semanales, mensuales, trimestrales o anuales. La metodología de Box y Jenkins busca encontrar un modelo matemático capaz de predecir basándose en el análisis del comportamiento de los datos en la serie temporal. A diferencia de muchos métodos, esta técnica no se basa en un patrón específico. En su lugar, utiliza un proceso iterativo para identificar un modelo adecuado a partir de modelos más generales.

La elección del modelo se somete a verificación con los datos, evaluando si logra describir la serie temporal con precisión. La validez del modelo se establece si los residuos entre las predicciones del modelo y los datos son pequeños y presentan una distribución independiente y aleatoria. En caso contrario, si el modelo no se ajusta o no proporciona predicciones fiables, se repite el proceso, explorando alternativas para encontrar un modelo que ofrezca una precisión satisfactoria (Hanke y Reitsh, 2014).

3.2.4. Modelos ARIMA

conocidos como modelos de autorregresión integrada de promedio móvil, forman una categoría dentro del enfoque general de Box y Jenkins para el análisis de series de tiempo estacionarias. Una serie se considera estacionaria cuando su media permanece constante, lo que significa que no cambia con el tiempo. Los modelos

AR y MA son parte de esta categoría ya que incorporan términos autorregresivos y de promedio móvil respectivamente. La metodología de Box y Jenkins permite seleccionar el modelo que mejor se adapta a los datos. Mediante esta técnica, se emplean enfoques autorregresivos y de promedio móvil para abordar cuestiones de pronóstico en el contexto de series temporales (Hanke y Reitsh, 2014).

Modelo Autorregresivo (AR): son aquellos que reflejan su propia naturaleza en el análisis. Esto implica que la variable que se estudia como dependiente y la variable que actúa como explicativa son en esencia la misma, pero con la particularidad de que la variable dependiente se posiciona en un momento temporal posterior (t) con respecto a la variable independiente (t-1) (Hanke y Reitsh, 2014).

Tiene la forma:

$$Y_t = \emptyset_0 + \emptyset_1 Y_{t-1} + \emptyset_2 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + \varepsilon_t$$

Donde:

Y_t : Variable dependiente del tiempo

$Y_{t-1}, Y_{t-2}, \cdots, Y_{t-p}$: Variables independientes que son variables dependientes desfasadas un número específico de periodos.

$\emptyset_0, \emptyset_1, \emptyset_2, \cdots, \emptyset_p$: Coeficientes que serán estimados.

ε_t: Residuo que representa sucesos aleatorios no explicados por el modelo.

Modelo de Promedio Móvil (MA): Este enfoque representa una aproximación para el análisis de series temporales univariados. El modelo de promedio móvil establece que la variable de interés se relaciona de manera lineal con su valor actual, así como con múltiples de sus valores previos, en presencia de un componente estocástico (Hanke y Reitsh, 2014).

Tiene la forma:

$$Y_t = w_0 + \varepsilon_t - W_1 \varepsilon_{t-1} - W_2 \varepsilon_{t-2} - \cdots - W_q \varepsilon_{t-q}$$

Donde:

Y_t: Variable dependiente.

$w_0, w_1, w_2, \cdots, w_q$: Coeficientes.

ε_t: Residuo o error.

$\varepsilon_{t-1}, \varepsilon_{t-2,\cdots,}\varepsilon_{t-q}$: Valores previos de residuos.

3.2.5. Prueba de Normalidad

Las pruebas de normalidad son herramientas estadísticas utilizadas para examinar si un conjunto de datos sigue una distribución normal. En otras palabras, estas pruebas evalúan si los datos se ajustan a la forma característica de una distribución gaussiana o de campana de Gauss, donde la mayoría de los valores se encuentran cerca de la media y hay una simetría en torno a esta. Las pruebas de normalidad son importantes para verificar si los datos exhiben esta propiedad, ya que muchas técnicas estadísticas asumen que los datos están distribuidos normalmente para que los resultados sean válidos y confiables (Carmona, M., & Carrión, H.,2015).

Estas pruebas permiten determinar si hay desviaciones significativas de la normalidad en los datos. Al realizar una prueba de normalidad, se calcula un valor estadístico basado en las diferencias entre la distribución de los datos y la distribución normal teórica. Este valor se compara con un umbral o valor crítico para tomar una decisión sobre si se debe rechazar o no la hipótesis de que los datos siguen una distribución normal. Si el valor calculado es mayor que el valor crítico, puede indicar que los datos no siguen una distribución normal (Carmona, M., & Carrión, H.,2015).

H0: Los datos siguen una distribución normal.

H1: Los datos no siguen una distribución normal.

3.2.6. Kolgómorov – Smirnov (KS)

La prueba de Kolmogorov-Smirnov es una técnica estadística utilizada para determinar si un conjunto de datos se ajusta a una distribución específica, generalmente la distribución normal. Esta prueba evalúa qué tan bien coinciden los datos observados con la distribución teórica, permitiendo medir si existen desviaciones significativas de la forma esperada de la distribución (Pedrosa et al. 2015).

El objetivo principal de esta prueba es ayudar a los investigadores a tomar decisiones informadas sobre si los datos siguen una distribución particular. Esto es relevante para seleccionar las técnicas estadísticas adecuadas y comprender los resultados. Sin embargo, es esencial utilizar esta prueba junto con otros enfoques y considerar el contexto y las limitaciones de los datos, ya que las conclusiones pueden variar según el tamaño de la muestra y otros factores (Carmona y Carrión,2015).

Hipótesis:

H0: Los datos siguen una distribución normal.

H1: Los datos no siguen una distribución normal.

Estadístico de contraste

$$D = Max_i \left| F(X_i) - \frac{i}{n} \right|$$

Donde:

D: estadístico de prueba.

F(Xi): función de distribución acumulada empírica evaluada en el i-ésimo valor.

n: muestra.

3.2.7. Machine Learning

Aprendizaje automatizado o automático es conocido también como Machine Learning y viene a ser un subcampo para la ciencia de la computación, como también en la inteligencia artificial, su finalidad es desarrollar métodos para que los ordenadores operen procesos de aprendizaje basados en algoritmos en conjuntos de datos. Estos procesos deben ser autónomos para que sigan funcionando a pesar de que la base de datos haya cambiado. (Mustafa et al.,2012).

Freitas (2002) nos menciona que Machine Learning está dividido en dos encuadres, el aprendizaje supervisado y el aprendizaje no supervisado. Por otra parte, este aprendizaje contiene referencias de métodos estadísticos, pero su metodología no sigue procedimientos estadísticos, es decir, los métodos no se basan en realizar los supuestos estadísticos satisfactorios para llegar a una solución y por el contrario se basan en la misma información que contiene la respectiva base de datos.

Dataset o conjuntos de datos estructurados están caracterizados por estar de forma tabuladas en las columnas y filas, haciendo de forma fácil su manejo. Otra de sus características es tener una variable aleatoria, esta variable representa 1 columna y cada una de las filas a una observación (Cabrera, 2018).

3.2.8. Aprendizaje Supervisado

Conocido también como aprendizaje automático supervisado y que pertenece a la rama de Machine Learning. Se determina como el uno de un conjunto de datos etiquetados para entrenar un algoritmo que va a clasificar o predecir con precisión los resultados esperados. una vez que el modelo ha sido alimentado con datos de entrada, el modelo adapta sus pesos hasta que esta se encuentre ajustada de forma

correcta. Este tipo de aprendizaje ayuda a dar soluciones a una diversidad de problemas de la vida real que se encuentran presentes en el mundo, así como realizar la clasificación de un correo no deseado para una carpeta que se encuentra separada del buzón de entrada (Hurwitz y Kirsch , 2018).

3.2.9. Aprendizaje no Supervisado

El aprendizaje automático no supervisado, emplea ciertos algoritmos para realizar análisis y agrupamiento de la información no etiquetada, de la misma manera agrupa los datos sin tener la necesidad de la intervención humana y detecta patrones que se encuentran ocultos. Cuenta con la gran capacidad de detectar semejanzas y divergencia en la información, lo cual se convierte en la perfecta solución para el desarrollo en los análisis exploratorios de los datos, planeamiento para realizar Cross Selling y segmentación de clientes (Schulman y Wolski, 2017).

3.2.10. Correlación

El estudio de la correlación se encuentra estrechamente relacionado con el modelo estadístico de regresión, pero se sabe que presentan una definición diferente. El análisis de correlación tiene como objetivo primordial medir el grado de relación lineal que tienen dos variables. Por ejemplo, encontrar la relación (coeficiente) entre las variables resultados de una evaluación de estadística y matemáticas. Por otro lado, el modelo de regresión nos estima o predice el resultado medio de la variable en función de valores fijos de las demás variables. Por lo tanto, puede ser una buena idea predecir el puntaje promedio en una evaluación de estadística en función de los puntajes de las evaluaciones de matemáticas de los estudiantes (Laguna,2014).

Hay una diferencia importante entre la regresión y la correlación la cual es importante alegar. Al realizar un análisis de regresión nos damos cuenta de que existe

cierta asimetría durante el procesamiento de las variables de interés. Consideran que la variable dependiente resulta ser estocástica o aleatoria, en otras palabras, tiene una distribución de probabilidad (Spìegel,1998).

Retornando al ejemplo, podemos decir que la correlación entre la valoración de las evaluaciones de los cursos de matemáticas y estadística vienen a ser las mismas que la correlación entre las puntuaciones de las evaluaciones de estadística y matemáticas. Los dos cursos o también conocidos como variables son consideradas de forma aleatoria. En conclusión, la teoría de correlación se basa la aleatoriedad de las variables, mientras que la teoría de regresión (la mayor parte) se basa en que la variable dependiente suele ser aleatoria, por lo que se refiere a las variables explicativas podemos decir que son fijas o aleatorias (Gujarati y Porter ,2010).

3.2.11. Regresión Lineal

Según Gujarati y Porter (2010), nos mencionan que dicho modelo de regresión presenta dos variables la cual está compuesta por una dependiente y otra independiente, por lo que estás se encuentran relacionadas o no, por otra parte menciona que dicho modelo trata de examinar la dependencia en la variable de estudio, también es conocida como la variable dependiente que presenta una relación con la variable explicativa cuya finalidad es estimar o predecir el valor medio de la población en la primera variable respecto a los valores conocidos.

Asimismo, podemos decir que existe otros modelos de regresión, la cual está compuesta por una variable dependientes y más de una variable explicativa, se conoce como modelo de regresión lineal múltiple.

3.2.12. Regresión Lineal Múltiple

Este modelo estadístico está conformado por una variable dependiente y más de una variable explicativa o independientes, dicho estudio nos ayuda a entender las causas

de las variaciones en nuestra variable de interés, es decir la variable dependiente. El modelo de regresión lineal múltiple nos ayudará a realizar el control de diversos tipos de situaciones de acuerdo a los estudios de interés que se realicen (Gujarati y Porter,2010).

Ecuación de regresión múltiple:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

Donde;

Y_i: i-ésimo valor de la variable dependiente.

X_{ik}: i-ésimo valor variables explicativas.

ε_i: i-ésimo error, perturbación estocástica.

El modelo en su forma matricial:

El modelo estadístico de regresión lineal múltiple tiene la expresión matricial de la siguiente forma.

$$Y = X\beta + u$$

Dicho modelo puede expresarse de forma matricial:

$$\hat{Y} = X\hat{B}$$

El modelo de la matriz y está representada de esta forma:

$$\vec{y} = [y_1\ y_2\ \dots\ y_n\ ]$$

La matriz X está representada así:

$$X = \begin{bmatrix} 1x_{11} & x_{12} & \dots & x_{1k} \\ 1x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

El vector, representa a los coeficientes y tiene esta forma:

$$\vec{\beta} = [\beta_0\ \beta_1\ \dots\ \beta_k\ ]$$

El siguiente vector es el error de la predicción.

$$\vec{e} = [e_0\ e_1\ \dots\ e_k\ ]$$

Calculando por el ajuste de mínimos cuadrados, tenemos:

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y$$

3.2.13. Prueba de hipótesis para una regresión lineal múltiple

Después de realizar la estimación de los parámetros del modelo de estudio nos preguntamos: ¿Cuáles son las variables independientes más importantes o significativas para realizar nuestro modelo? Luego realizamos una prueba de significación.

Prueba de significancia

Esta prueba nos permite determinar la existencia de alguna relación lineal entre la variable dependiente "y" y variables independientes "x". Las hipótesis se plantean de la siguiente forma:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 = al\ menos\ un\ B_j \neq 0$$

Se rechaza la hipótesis nula, cuando al menos una de las variables independientes $(X_1, X_2, \ldots, X_k)$ aporta al modelo de manera significativa.

Este proceso se resume en la siguiente tabla de análisis de la varianza o conocida también como ANOVA.

Tabla 1

Tabla ANOVA

Fuente de Variación	**Grado de Libertad**	**Suma de Cuadrados**	**Cuadrado Medio**	**F**
regresión	p-1	SCReg	CMReg	$F = \frac{CMReg}{CME}$
Error	n-p	SCE	CME	-
Total	n-1	SCT	-	-

Nota: Expresiones para el cálculo de la tabla ANOVA

La tabla 1 nos muestra la comparación de las varianzas para establecer si existe o no alguna diferencia significativa entre los otros grupos (Gujarati y Porter ,2010).

Donde:

$$SCT = \ y'y - n\underline{y}^2 \ \ldots..*$$

$$SCReg = \ \widehat{\beta}'x'y - n\underline{y}^2 \ \ldots..**$$

Reemplazando en la ecuación (*) y (**):

$$SCE = SCT - SCReg = \ y'y - \widehat{\beta}'x'y$$

Donde:

n: Número de datos al ser procesados.

P: Grados de libertad.

Prueba de coeficiente individual de una regresión:

A fin de elaborar la prueba de significancia de algún coeficiente individual, se desarrolla de la siguiente forma.

$$H_0: \beta_j = 0 \ (la\ variable\ x_j\ no\ aporta\ información\ al\ modelo).$$
$$H_1: \beta_j \neq 0 \ (la\ variable\ x_j\ aporta\ información\ al\ modelo).$$

El estadístico de prueba para realizar la siguiente hipótesis es:

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$

La hipótesis nula se rechaza si $|t_0| > t_{n-p\,;\,\propto\frac{1}{2}}$, de no rechazarse la hipótesis nula, se puede eliminar el regresor x_j del modelo.

IV. Metodología

Esta investigación se realizará con la orientación de la metodología CRISP-DM, los siguientes pasos se presentan a continuación.

4.1. Entendimiento del negocio

La empresa encargada de ofrecer servicios de gestión humana cuenta con todos los registros detallados de cada uno de los reembolsos aprobados ante Essalud.

Es por ello que el objetivo de la empresa es realizar la mayor cantidad de reembolsos frente a la gran demanda de solicitudes enviadas por cada cliente. El enfoque va dirigido a cada una de las personas encargadas de realizar el seguimiento de cada solicitud ingresada a Essalud, que en este caso serían las gestoras a cargo. Para alcanzar nuestro objetivo es importante realizar estudios y apoyarnos de herramientas y modelos estadísticos. El modelo de regresión lineal múltiple nos ayudará a conocer las variables que nos brinde una mejor gestión y por ende realizar nuevas estrategias.

4.2. Entendimiento de la data

Cada uno de los registros son almacenados en una base de datos la cual fue implementada por la empresa, con el fin de mantener su seguridad e integridad. Estos registros corresponden a los descansos médicos recibidos por incapacidad temporal de cada colaborador perteneciente a los clientes en el periodo de enero del 2019 hasta mayo del 2023.

Este periodo nos será de gran ayuda para poder realizar dicho estudio, sin embargo, no es necesario utilizar más registros debido a que la información ingresa día tras día y se cuenta con la información necesaria. Ahora mostramos algunas estadísticas de nuestras variables en la siguiente tabla.

Tabla 2

Estadísticos descriptivos de las variables

	Edad	**Días**	**Observaciones**	**Monto presentado**	**Monto reconocido por Essalud**	**Monto no reconocido por Essalud**
Media	43.45	24.47	1.14	2637.30	2630.07	7.23
Desviación estándar	9.10	24.01	0.84	3203.19	3166.52	1481.95
Cuartil 25	36	9	0	1051.38	1036.84	0
Cuartil 50	43	20	1	1831	1824	0
Cuartil 75	51	30	2	3130.50	3161.10	0

Nota: tabla descriptiva de las variables de interés para dicho estudio.

4.3. Preparación de los datos

Con la información obtenida se llevó a cabo un mapeo general para detectar incongruencias, se realizó la limpieza de los registros dado que se encontró registros duplicados e información incompleta, por ejemplo: se encontró registros que figuraban como enfermedad en su tipo de licencia y que corresponden a una maternidad, los clientes envían la información y al ser automatizada la actualización de los registros ingresantes, se realiza el mapeo respectivo para la validación de cada descanso médico, es así como se detectó un registro de una colaboradora de sexo femenino con 98 días de descanso médico, que figuraba con el tipo de licencia, enfermedad. Por otra parte, eliminamos los registros duplicados.

Cada una de las observaciones encontradas, fueron corregidas antes de realizar el análisis para la elaboración de dicho estudio, teniendo en cuenta el uso de todas nuestras variables de interés.

Las variables utilizadas para dicha investigación so

Tabla 3

Descripción de las variables

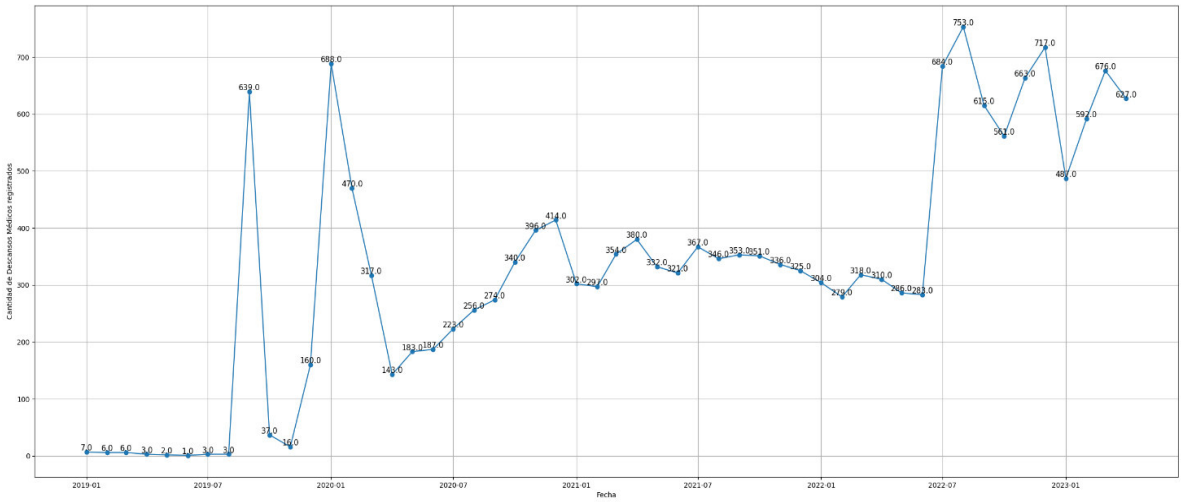
Variable	Definición	Tipo de Variable
Monto presentado	Monto total en soles que se solicita a Essalud por alguna incapacidad temporal.	Cuantitativa Continua
Monto reconocido ante Essalud	Monto total en soles, reconocido por Essalud.	Cuantitativa Continua
Monto no reconocido ante Essalud	Monto total en soles, no reconocido por Essalud.	Cuantitativa Continua
Edad	Cantidad de años vividos en una persona.	Cuantitativa Discreta
Días	Número de días de incapacidad temporal.	Cuantitativa Discreta
Observaciones	Cantidad de veces en la que se observó una solicitud de reembolso	Cuantitativa Discreta

Nota: Definición y tipos de variables de interés.

Antes de elaborar el modelo de estudio, verificamos el grado de correlación entre las variables, para tener claro al momento de trabajar con las variables de interés y no exista inconsecuencia.

Figura 3

Cantidad de días en los descansos médicos registrados durante el periodo enero 2019 -abril 2023

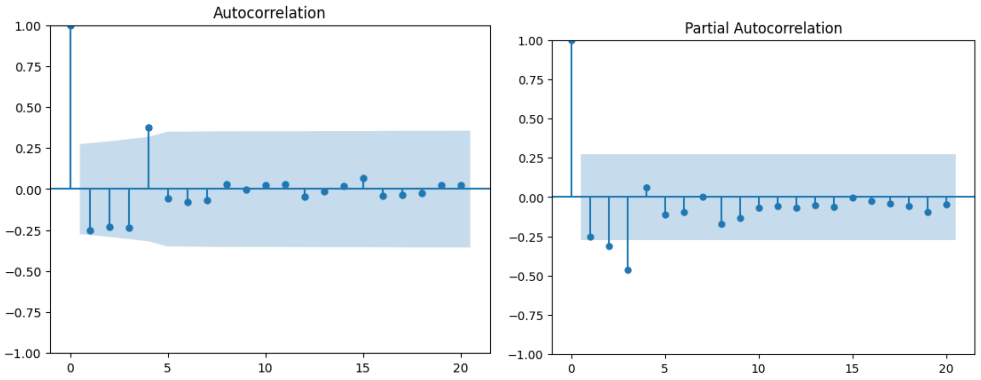


Nota: resultado de la serie de tiempo de la cantidad de descansos médicos.

En la figura 3, se observa las cantidades de días registrados en los descansos médicos que corresponden a cada mes. En el mes de septiembre del 2019 se registró un aumento significativo en la cantidad de personas con un total de 639. A inicios del año 2020 empezó a mostrar una cantidad de personas con descansos médicos a comparación del año anterior, por otro lado, en agosto del 2022 se mostró 753 registros; siendo ese el caso con mayores descansos médicos en todo el periodo establecido.

Figura 4

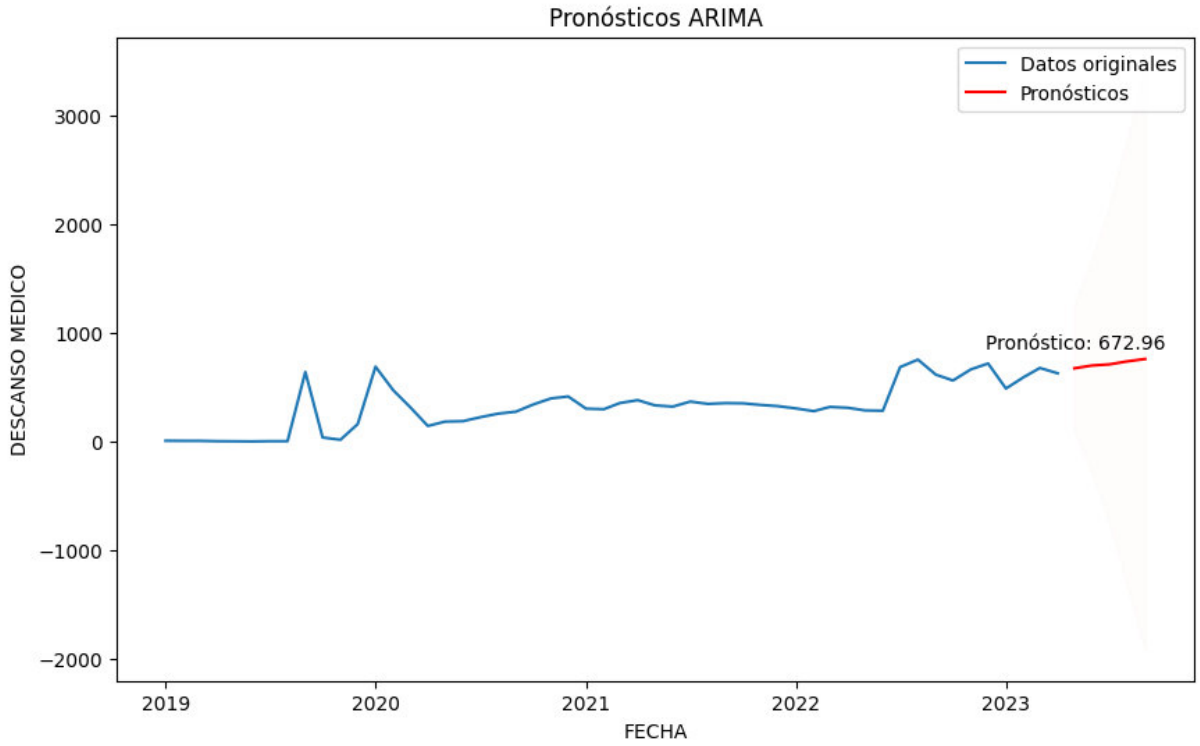
Función de autocorrelación



En la figura 4 observamos valores por encima del área sombreada, es decir estos valores son significativamente distintas del valor cero.

Figura 5

Pronósticos de la cantidad de días en los descansos médicos para mayo del 2023

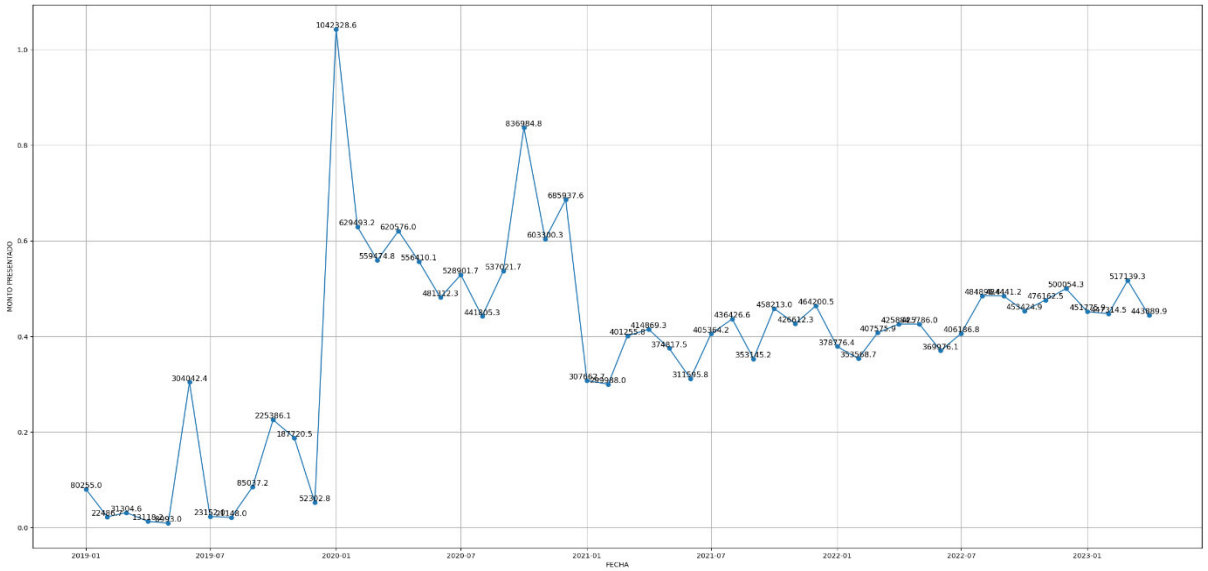


Nota: Pronostico de la cantidad de descansos médicos para el mes de Mayo.

Realizando el análisis respectivo se obtuvo el pronóstico de la cantidad de días de todos los descansos médicos registrados para el mes de mayo del 2023 con un modelo ARIMA (2,3,1) y cuyo valor pronosticado muestra una ligera tendencia ascendente en el próximo mes con un total de 673 registros que se deberán de presentar a Essalud para el reembolso respectivo.

Figura 6

Montos presentados para el reembolso en el periodo enero 2019 -abril 2023

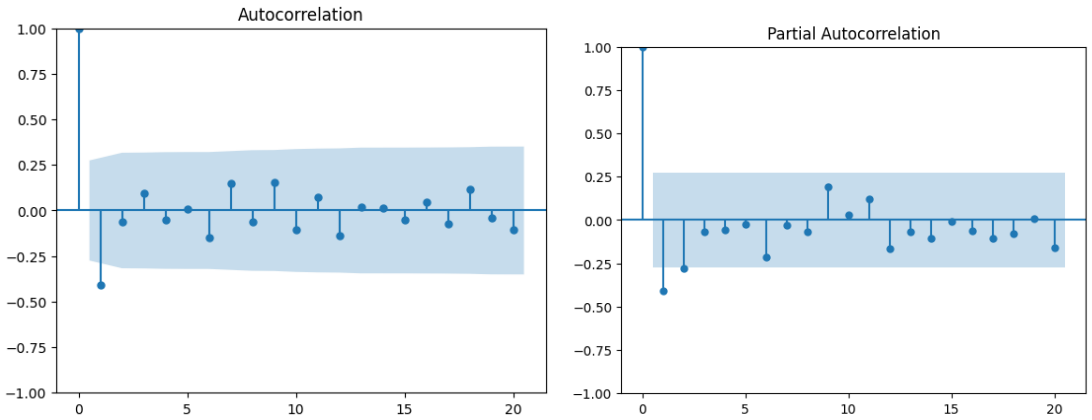


Nota: resultado de la serie de tiempo de los montos presentados ante Essalud.

En la figura 6, observamos que el pico más alto que se presentó en todo el periodo fue en enero del 2020 con un total de S/.1,042,328.6, también se refleja que a partir de enero del 2021 se presentaron montos representativos que se mantienen en un rango de S/.307,662.2 a S/.517,139.3 este aumento nos hace referente al contexto de la covid-19.

Figura 7

Función de autocorrelación

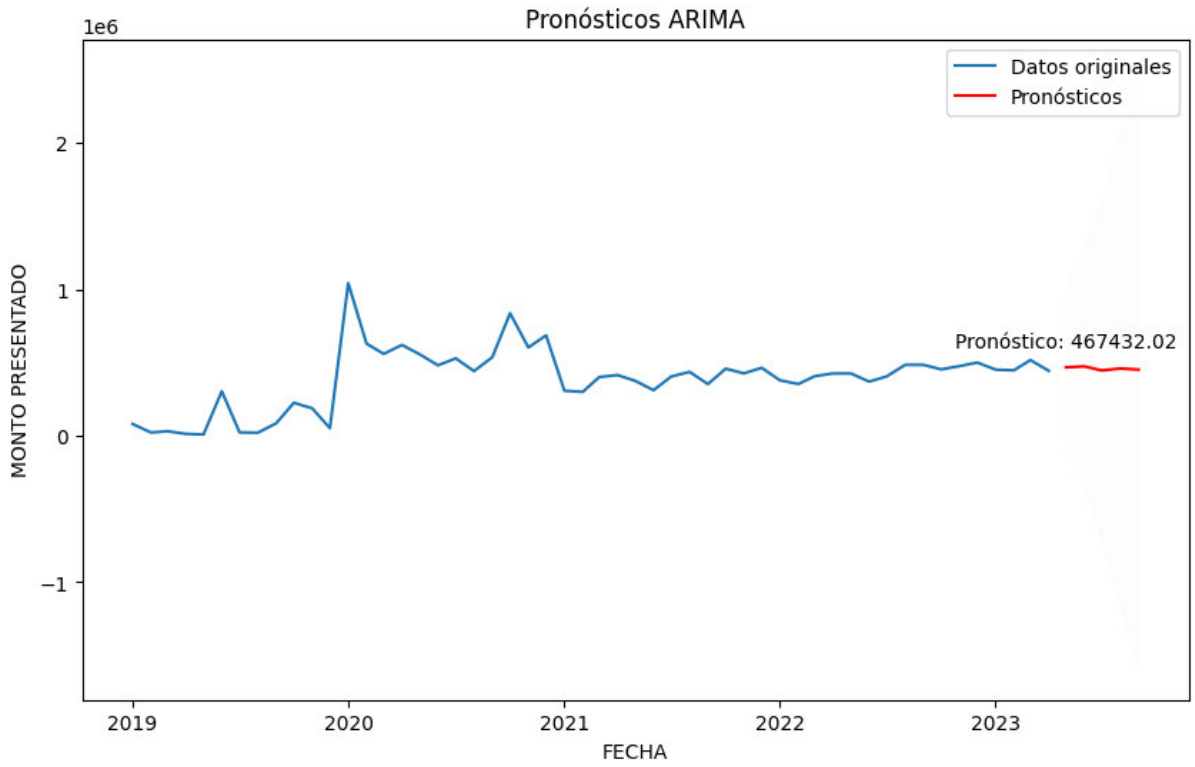


En la figura 7, analizando conjuntamente la gráfica que nos brinda los resultados del AR y Ma podemos decir que se observan valores ligeramente significativos que se

encuentran fuera del área sombreada.

Figura 8

Pronósticos de los montos presentados para mayo del 2023

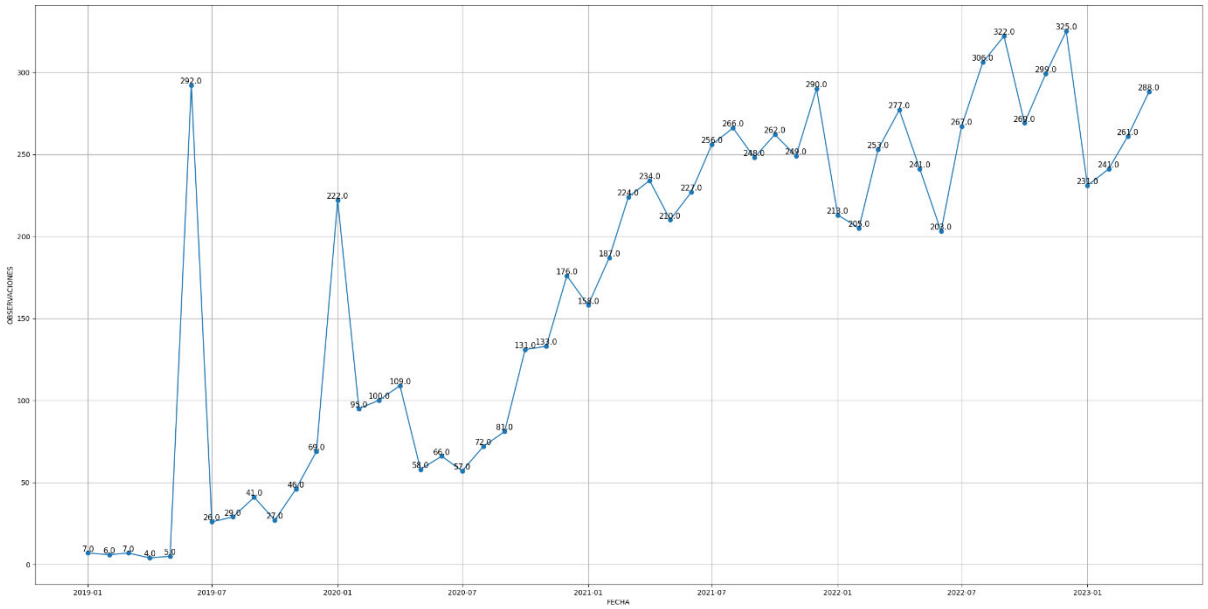


Nota: Pronostico del monto a presentar para el mes de Mayo.

En el grafico 8 nos muestra el pronóstico de los montos presentados a Essalud hasta el mes de abril del 2023, el mejor modelo para obtener el monto que se debe presentar en el mes de mayo es el modelo ARIMA (2,2,0) y la cual nos brinda un monto de S/.467,432.02.

Figura 9

Observaciones obtenidas ante la solicitud de reembolso en el periodo enero 2019 -abril 2023

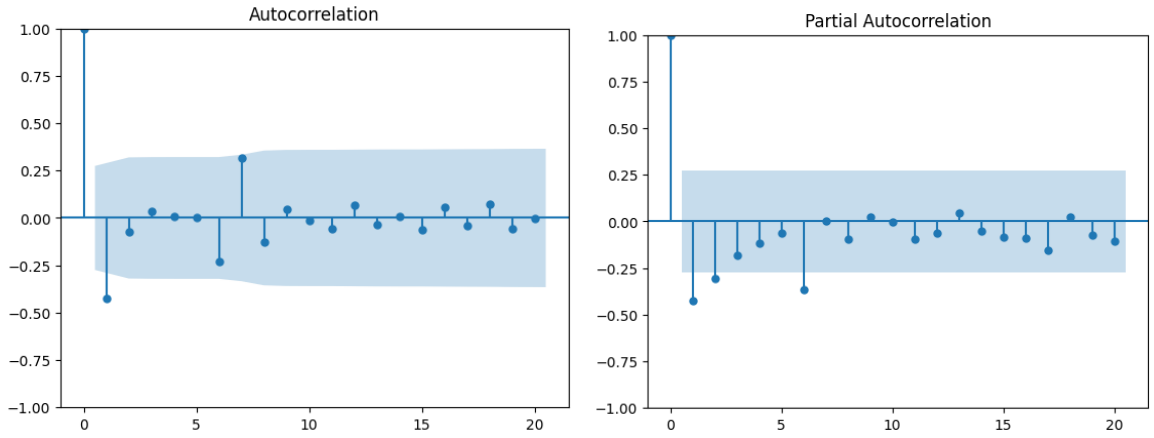


Nota: resultados de la serie de tiempo de las observaciones obtenidas.

En la figura 9, Podemos decir que las observaciones hechas por Essalud ante el ingreso de las solicitudes de reembolso van en aumento. En diciembre del 2022 se mostró un registro de 325 observaciones, siendo este el registro más alto.

Figura 10

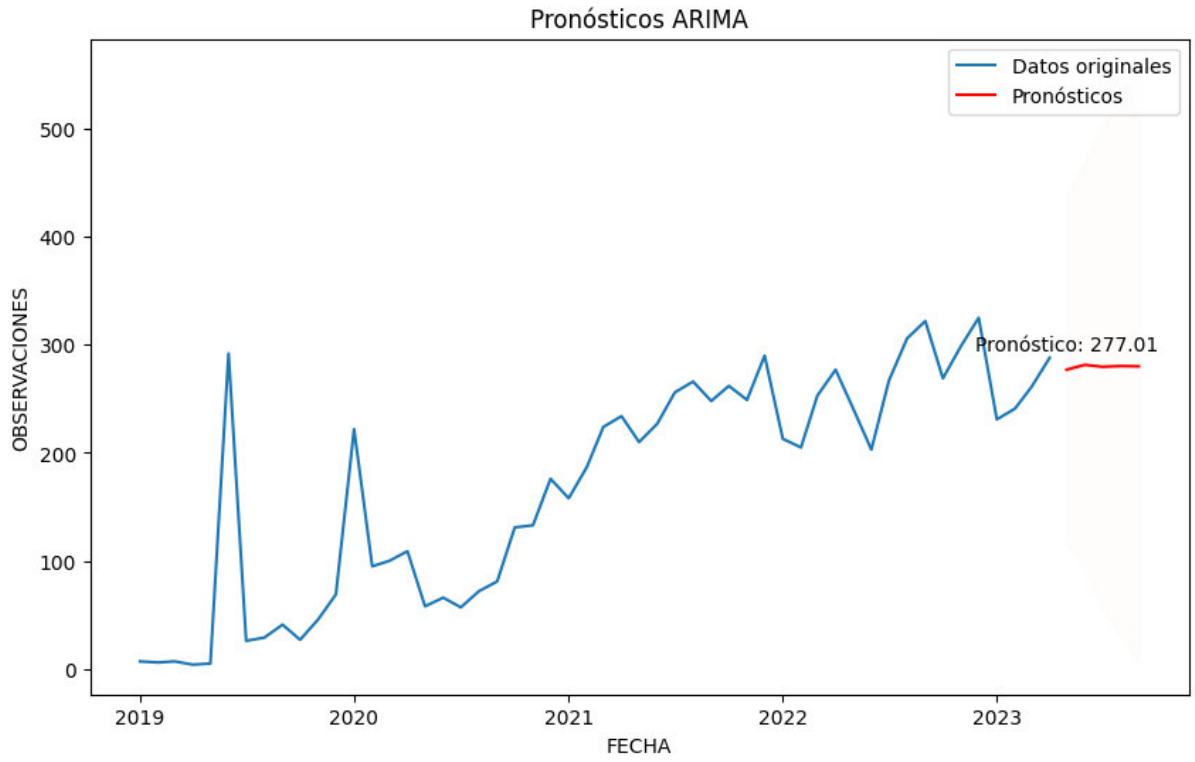
Función de autocorrelación



Siguiendo el enfoque propuesto por Box-Jenkins, en el análisis del ACF y PACF elegiremos al modelo adecuado, por lo que observamos valores ligeramente significativos es decir se encuentran fuera del área sombreada.

Figura 11

Pronósticos de las observaciones emitidas por Essalud para mayo del 2023

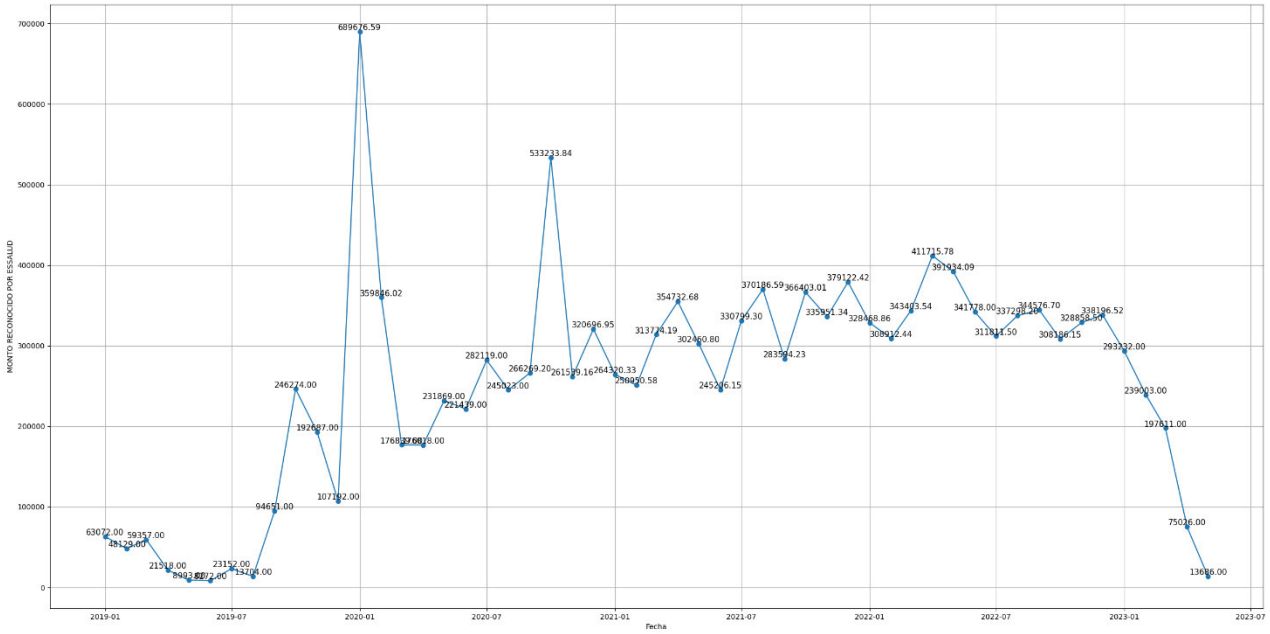


Nota: Pronóstico de las posibles observaciones emitidas por Essalud para el mes de Mayo.

De acuerdo con la gráfica, podemos decir que tenemos el pronóstico de las posibles observaciones emitidas por Essalud para mayo del presente año y cuyo modelo adecuado fue ARIMA (1,1,0); dicho modelo nos brinda 277 observaciones.

Figura 12

Montos reconocidos ante Essalud en el periodo enero 2019 -abril 2023



Nota: resultado de la serie de tiempo de los montos reconocidos.

De acuerdo con la figura 12 nos muestra los montos reconocidos de las solicitudes presentadas ante Essalud, el subsidio más alto que se logró recuperar fue de S/.689,676.00 perteneciente a enero del 2020. Se observa que los montos recuperados se mantuvieron en un rango de S/.359,846.00 a S/.338196.00 en el periodo de febrero 2020 y diciembre 2022 respectivamente, por lo que se observa una caída en los montos a recuperarse a partir de enero del 2023.

Tabla 4

Resultados

VARIABLE	VALOR PRONOSTICADOS	RMSE	R^2	MAE	MODELO
Días	673	0.43	0.83	0.56	ARIMA (2,3,1)
Montos presentados	S/.467,432.02	0.57	0.79	0.43	ARIMA (2,2,0)
Observaciones	277	0.41	0.91	0.33	ARIMA (1,1,0)

Nota: resumen de los resultados obtenidos en las variables mediante el modelo ARIMA.

De acuerdo con la tabla 4 podemos observar los resultados de las variables de estudio en el pronóstico realizado. Se observa que las variables días, monto presentado y observaciones tienen como modelo al ARIMA (2,3,1), ARIMA (2,2,0) y ARIMA (1,1,0) respectivamente, dichos modelos son adecuados debido a que su RMSE fueron los valores más bajos.

Tabla 5

Prueba de normalidad

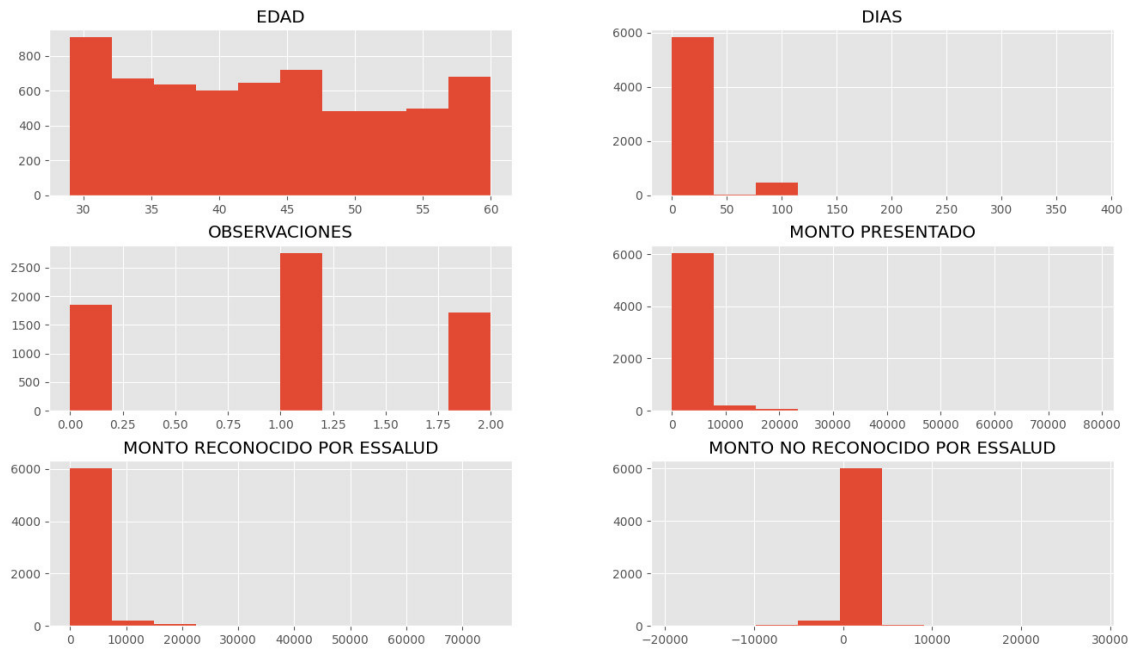
VARIABLE	P_VALOR	H0	H1
Edad	0.12	Los datos analizados siguen una distribución normal	Los datos analizados no siguen una distribución normal
Días	0.23		
Observaciones	0.01		
Monto presentado	0.05		
Monto reconocido ante Essalud	0.15		
Monto no reconocido ante Essalud	0.00		

Nota: resultado de la prueba de normalidad, Kolmogorov- smirnov (KS).

En la tabla se observa que el p_ valor para las variables Edad, Días, Observaciones, Monto presentado, Monto reconocido ante Essalud y Monto no reconocido ante Essalud es menor al nivel de significancia (α= 0.05). Se rechaza la hipótesis nula, es decir que con un nivel de significancia del 5% existe evidencia estadística suficiente para afirmar que las variables Edad, Días, Observaciones, Monto presentado, Monto reconocido ante Essalud y Monto no reconocido ante Essalud no siguen una distribución normal.

Figura 13

Histogramas de las variables



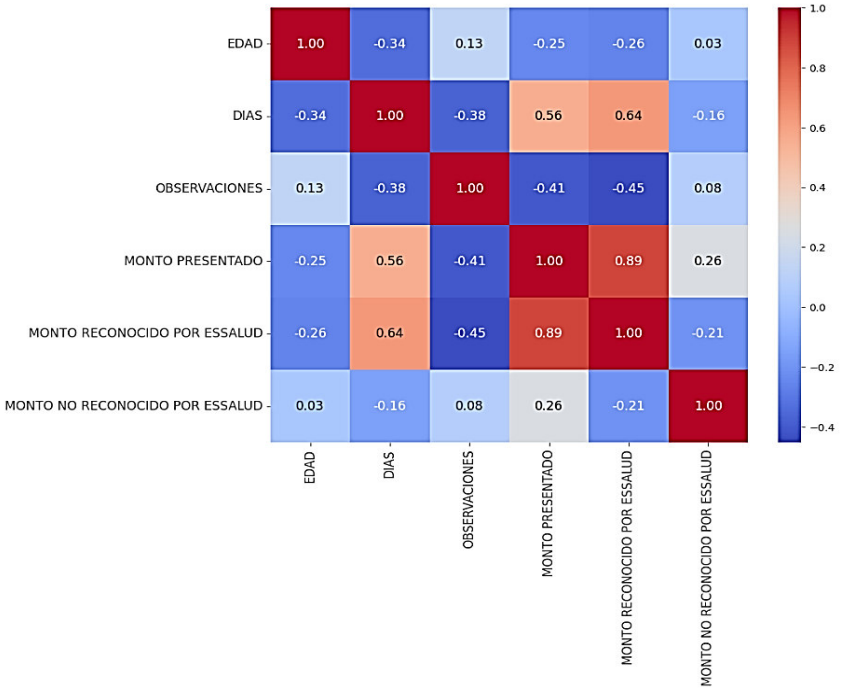
Nota: resultados del análisis descriptivo de cada una de las variables de estudio.

En la Figura observamos la información de cada variable, es decir en la variable edad se muestra una gran cantidad de registros de aquellas personas que se encuentran entre 22 a 32 años y que contaron con al menos un descanso médico, Por otro lado, se observa que de las 16 994 solicitudes presentadas; más de 25 mil registros contaron con al menos una observación al ingresar la solicitud de reembolso.

Aparentemente el monto reconocido por Essalud y el monto presentado por la empresa vienen a ser similares.

Figura 14

Matriz de Correlaciones



Nota: correlación entre las variables de estudio.

La siguiente matriz de correlación observamos que existe una moderada y alta correlación, cuyo valor mínimo y máxima es de -0.45 y 0.89 respectivamente, por lo que podemos decir que el valor de una variable puede proporcionar información valiosa para poder predecir el valor de la variable respuesta.

4.4. Modelado y validación del modelo de regresión

Monto reconocido por

Essalud = 185,346.6+6,965*dias+261,383*observaciones+0.390*monto presentado

Coeficiente de determinación (R^2): 0,8565

Coeficiente de correlación múltiple: 0,9254

Monto reconocido por

Essalud = 185,346.6+7*dias+261*observaciones+0.39*monto presentado

Monto reconocido por

Essalud = 185,346.6+7*673+261*277+0.39*467132.02

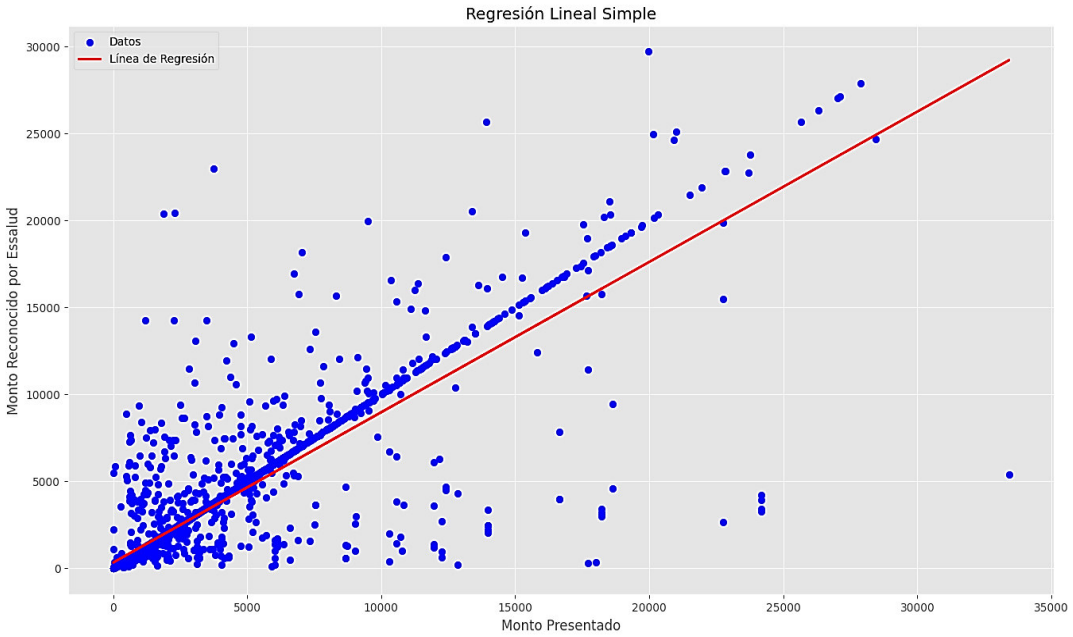
Monto reconocido por

Essalud =444,653.088

El modelo nos indica que por cada unidad del monto reconocido por Essalud, los días aumentan aproximadamente en 7 días, mientras que el número de observaciones aumenta en 261 unidades y el monto presentado aumenta en 0.39. Observando el coeficiente de determinación podemos afirmar que las variables independientes días, observaciones y monto presentado; explican con un 85% a la variable dependiente monto reconocido por Essalud.

Figura 2

Gráfico de dispersión



Nota: correlación $r > 0$.

En la figura 15 nos muestra una correlación lineal positiva, entre las variables Monto presentado y el Monto reconocido por Essalud.

V. Conclusiones

Sin duda, al realizar el estudio de recupero de subsidios ante Essalud viene a ser un tema que se debe tener análisis y entendimiento de la base de datos, si bien es cierto que cada empresa cuenta con colaboradores de distintas edades y diferente genero; es primordial conocer la población que conforma estas entidades.

Al realizar el análisis de tiempo mediante el modelo ARIMA de las variables de estudio podemos concluir que el mejor modelo para pronosticar la cantidad de días de descansos médicos que se presentaran para el mes de mayo fue ARIMA (2,3,1) con un RMSE = 0.43, MAE = 0.56, R^2= 0.83 y un pronóstico de 673 días. La variable observaciones tuvo como modelo adecuado a ARIMA (1,1,0) con un RSME = 0.41, MAE = 0.33, R^2= 0.91 y un pronóstico de 277 observaciones por parte de Essalud para el mes de mayo por otro lado podemos decir que el monto por presentar para el mes de mayo es de S/.467,432.02 este valor fue pronosticado por el modelo ARIMA (2,2,0) con un RSME = 0.57, MAE = 0.43 y R^2= 0.57.

Como se puede observar, estos valores pronosticados nos ayudaron a explicar cómo se puede trabajar de forma adecuada al momento de realizar un subsidio ante Essalud. Dichos valores fueron reemplazados en el modelo apropiado de regresión lineal múltiple, este modelo tiene la forma de: Monto reconocido por Essalud = 185,346.6 + 7*días + 261*observaciones + 0.39*monto presentado, con un R^2= 0.8565. Al realizar el reemplazo de los valores nos brindo un monto reconocido de S/. 444,653.088 tal y como se muestra en la ecuación, las variables que nos permiten proyectar una buena gestión de subsidios son la cantidad de días en un descanso médico, también tener en cuenta la cantidad de observaciones que nos va a emitir Essalud y el monto que se presentará.

Para concluir con dicha investigación, obtenemos como resultado una correlación lineal positiva de 0.89, entre las variables Monto presentado y el Monto reconocido por Essalud.

VI. Recomendaciones

Una vez concluida la investigación se recomienda a la empresa, brindar capacitación a todas las personas involucradas en realizar los trámites y seguimientos de los subsidios sobre los papeleos correspondientes a las solicitudes de reembolso antes Essalud, cuya finalidad es evitar tener observaciones por parte de Essalud y así poder generar mayor efectivad de la que ya se tiene en la gestión de reembolsos antes Essalud.
Por otro lado, se recomienda realizar una eficiente elaboración de cálculos al momento de generar el monto que se debe presentar, ya que puede ser un indicador que interfiera en la meta establecida.
Se usó el modelo de Regresión Lineal Múltiple para estudiar las variables de interés, se sugiere realizar en un siguiente estudio un modelo en la cual permita estudiar a las variables cualitativas, tales como tipo de licencia (enfermedad o maternidad) y genero.

VII. Bibliografía

Angelucci-Bastidas L, Rondón-Bernard JE. (2018). Adherencia al tratamiento en diabetes tipo 2: Un modelo de regresión logística. Caracas 2017-2018. MÉD.UIS.2021;34(2): 29- 39. http://www.scielo.org.co/pdf/muis/v34n2/1794-5240-muis-34-02-29.pdf

Banco central de Reserva del Perú [BCRP]. (2017). *Series estadísticas.* https://estadisticas.bcrp.gob.pe/estadisticas/series/.

Cabrera, D. (2018). *Modelado de sistemas dinámicos con Machine Learning: aplicaciones al mantenimiento basado en la condición.Tesis Doctoral*. Universidad de Sevilla, Sevilla

Campos, T. S. R. (2010). *Manual de seguridad social: Tratamiento de las prestaciones en salud y pensiones*. Lima: Gaceta Jurídica. p.21-22

Carmona, M., & Carrión, H. (2015). *Potencia de la prueba estadística de normalidad Jarque-Bera frente a las pruebas de Anderson-Darling, Jarque-Bera robusta, Chi cuadrada, Chen-Shapiro y Shapiro-Wilk* .Universidad Autónoma del Estado de México. https://core.ac.uk/download/pdf/159384191.pdf

Castillo, O. (2022). *Desarrollo de modelos predictivos de regresión en la industria minera mediante el uso de algoritmo de machine learning.* Universidad Nacional Mayor de San Marcos, Perú. https://cybertesis.unmsm.edu.pe/handle/20.500.12672/18458

Chakravarti, I.M., Laha, R.G., & Roy, J. (1967). Kolmogorov-Smirnov (K-S) test. En Handbook of Methods of Applied Statistics, Volume I (pp. 392394). New York: Wiley.

Cruz, A., Martinez, G., Martinez, A., Morales, I., Escamirosa, C. (2023). Estimación de las tendencias de precios del agave mezcalero en México utilizando modelos de regresión lineal múltiple, Mexico.

Freitas, A. (2002). *Data mining and knowledge discovery with evolutionary alghorithms.* Springer Science & Business Media.

Gujarati, D. y Porter, D. (2010). *Econometría,* México, México: McGRAW-

Hanke, J. y Reitsh, A. (2014). Pronósticos en los Negocios. https://cbtis177.edu.mx/pdf/biblioteca_virtual/admon_rec_humanos/Pronosticos_en_los_Negocios_Reitsch_5a_Ed.pdf

HILL/INTERAMERICANA EDITORES, S.A. (pag 15-21)

Hurwitz, J., y Kirsch, D. (2018). Augmented intelligence: the business power of human–machine collaboration. Auerbach Publications.

Laguna, C. (2014). Correlación y regresión lineal. Instituto Aragonés de Ciencias de la Salud, 4, 1-18.

Lizares, M. (2017). *Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico.* Universidad Nacional Mayor de San Marcos, Perú. https://cybertesis.unmsm.edu.pe/bitstream/handle/20.500.12672/7122/Lizares_cm.pdf?sequence=3&isAllowed=y

López-Herrera, N., Aceros, M. & Luzardo, M. (2019). *Análisis de los hurtos en Colombia durante el año 2017 mediante los modelos de regresión lineal múltiple y la regresión ponderada geográficamente*. Revista Criminalidad, 61(3): 141-163

Mustafa, M., Isa, R., y Rezaur, R. (2012). Artificial neural networks modeling in water resources engineering; infrastructure, and applications, Int. J. of Civil, Environmental, Structural, Construction and Architecture Engineering, 6(2) 128-136.

Organización de las naciones Unidas (1948) . Declaración Universal de los Derechos Humanos. París: Asamblea General de las Naciones Unidas.

Ortiz Cardona I., (2020). *propuesta de modelo arima para la serie temporal de los casos de covid-19 en Colombia aplicando la metodología box and Jenkins.* https://repository.libertadores.edu.co/handle/11371/3594

Pedrosa, I., Juarros-Basterretxea, J., Robles-Fernández, A., Basteiro, J., & García-

Cueto, E. (2015). Pruebas de bondad de ajuste en distribuciones simétricas,¿ qué estadístico utilizar?. *Universitas psychologica*, *14*(1), 245-254.

Quispe, C. (2018). *Implementación de un modelo predictivo para incrementar la captación de seguros en una entidad financiera.* Universidad Nacional Mayor de San Marcos, Perú. https://cybertesis.unmsm.edu.pe/bitstream/handle/20.500.12672/11715/Quispe_cc.pdf?sequence=1&isAllowed=y

Sanchez, C. (2022). *Proceso de análisis jerárquico y regresión lineal múltiple para la priorización de estrategias de negocio en una central de riesgos en Lima.* Universidad Nacional Mayor de San Marcos, Perú. https://cybertesis.unmsm.edu.pe/bitstream/handle/20.500.12672/18610/S%c3%a1nchez_bc.pdf?sequence=3&isAllowed=y

Schulman, J., y Wolski, F. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

SEGURO SOCIAL DE SALUD. (2012). DIRECTIVA N°08-GG-essalud-2012. *normas complementarias al reglamento de pago de prestaciones económicas*. Lima, Perú.

Sommerfeldt,T. (2020). *Inteligencia emocional y estrés laboral en docentes de educación escolar básica durante la pandemia covid- 19.* Universidad Nacional de Itapúa, Paraguay.

Spiegel, M. R. (1998). Estadística.