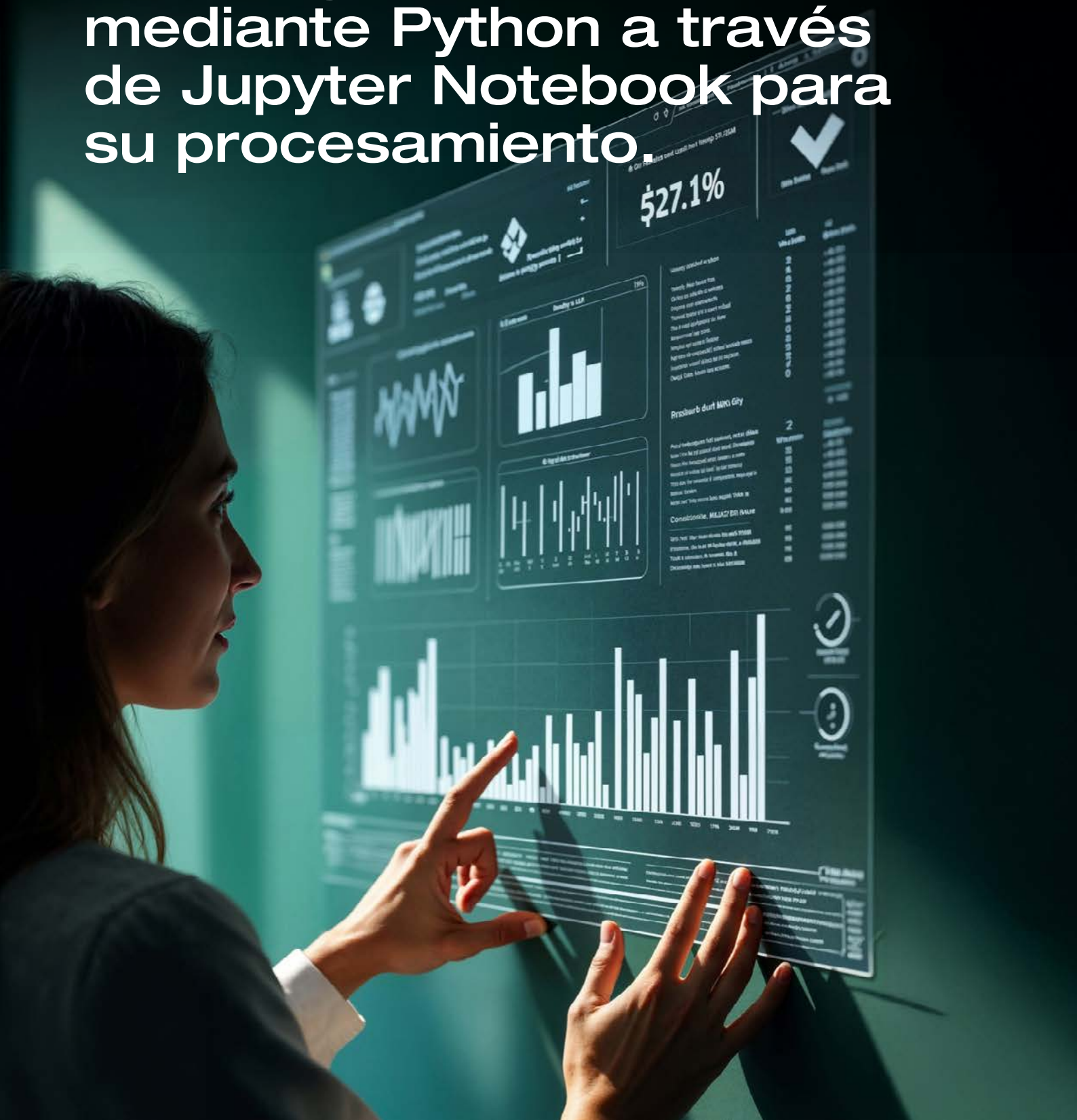
Construcción de un modelo

de series de tiempo de tipo forecasting autorregresivo recursivo, mediante Python a través de Jupyter Notebook para su procesamiento.



José Luis Espinoza Melgarejo, Universidad Nacional del Santa

Modelo de series de tiempo para predecir la demanda de atención de pacientes con enfermedad renal crónica, 2022

Time series model to predict the demand for care of patients with chronic kidney disease, 2022

RESUMEN

El objetivo principal de este trabajo es pronosticar la demanda de pacientes con enfermedad crónica renal en establecimientos de salud estatales del Perú en 2022 mediante modelos de series de tiempo y realizar un análisis descriptivo de dicha demanda. Este estudio se justifica, ya que no existen estudios de este tipo en Perú, aun sabiendo de las carencias en equipamiento e insumos para el tratamiento de enfermedades renales mediante procedimientos médicos como la diálisis.

Se trata de un estudio de alcance descriptivo y exploratorio; el diseño es no experimental, transversal y descriptivo. La población está conformada por 1 064 744 registros de pacientes con información variada como periodo de atención, código de identificación, nombre del establecimiento de salud, entre otros tomados de la Plataforma de Datos Abiertos del Perú. No se realizó un muestreo debido a que se construyeron modelos de series de tiempo en intervalos diarios. Se usaron técnicas estadísticas como gráficos de barras simples y apilados, gráficos circulares y tablas de frecuencias; se construyó un modelo de series de tiempo de tipo *forecasting* autorregresivo recursivo, mediante Python a través de Jupyter Notebook para su procesamiento.

Los resultados más importantes muestran que la mayor demanda se concentra en Lima, con una distribución equilibrada entre hombres y mujeres, y una mayor incidencia en personas de 50 a 70 años, especialmente entre quienes tienen seguro gratuito. Analizando las componentes de la serie de tiempo y haciendo uso de la prueba de Dicky-Fuller, se optó por emplear un modelo *forecasting* autorregresivo recursivo obteniendo con R^2 de 96,62 %. Además, luego de realizar un ajuste de hiperparámetros, se logró obtener un R^2 de 94,61 % para el mismo modelo, siendo este menos sobre ajustado y cumpliendo con la mayoría de los supuestos de series de tiempo.

Por tanto, podemos concluir que el modelo obtenido es bueno para predecir la demanda de pacientes demanda de atención de pacientes con enfermedad renal crónica, ya que tiene un desempeño óptimo y cumple con todos los supuestos a excepción de la autocorrelación.

ABSTRACT

The main objective of this study is to forecast the demand for patients with chronic kidney disease in state-run healthcare facilities in Peru in 2022 using time series models, and to conduct a descriptive analysis of this demand. This study is justified as there are no similar studies in Peru, despite the known deficiencies in equipment and supplies for the treatment of kidney diseases through medical procedures such as dialysis.

This is a descriptive and exploratory study; the design is non-experimental, cross-sectional and descriptive. The population consists of 1,064,744 patient records with various information, such as the care period, identification code, name of the healthcare facility, among others, taken from the Open Data Platform of Peru. No sampling was carried out because time series models were built at daily intervals. Statistical techniques such as simple and stacked bar graphs, pie charts and frequency tables were used; A recursive autoregressive forecasting time series model was built using Python through Jupyter Notebook for processing.

The most important results show that the highest demand is concentrated in Lima, with a balanced distribution between men and women, and a higher incidence in people aged 50 to 70, especially among those with free insurance. Analyzing the components of the time series and using the Dicky-Fuller test, it was decided to use a recursive autoregressive forecasting model, obtaining an R2 of 96.62%. In addition, after performing

Palabras Claves

Demanda de pacientes, enfermedad renal crónica, series de tiempo, forecasting autorregresivo, ajuste de hiperparámetros, supuestos de series de tiempo.

Key words

Patient demand, chronic kidney disease, time series, autoregressive forecasting, hyperparameter tuning, time series assumptions..

a hyperparameter adjustment, an R2 of 94.61% was obtained for the same model, which was less over-adjusted and met most of the time series assumptions [3].

Therefore, we can conclude that the model obtained is good for predicting the demand for care of patients with chronic kidney disease since it has an optimal performance and meets all the assumptions except for autocorrelation.

INTRODUCCIÓN

La enfermedad renal crónica (ERC) es un problema de salud pública con una alta mortalidad cardiovascular y elevados gastos en salud. Se estima un trillón de dólares en cuidados para pacientes con ERC a nivel mundial [9]. Se calcula que el 50 % de la población de pacientes con ERC no recibe terapia de reemplazo renal (TRR), y en algunas regiones del Perú no existen centros de diálisis del Ministerio de Salud (Minsa) para atenderlos, lo que implica la necesidad de duplicar el presupuesto de salud [9].

De acuerdo con Zanabria-Calderón (2022) [21], la carga de esta enfermedad se ha incrementado con 21 millones de nuevos casos de ERC en 2016 y 1,2 millones de muertes anuales. Además, EsSalud cuenta con estadísticas de demanda atendida de la ERC desde 2010 a 2016 [22].

A pesar de la alta mortalidad de la ERC, la investigación en Perú es limitada. Dado que los principales factores de riesgo varían según la región, se deduciría que la mortalidad no es uniforme en todas las regiones, siendo la sierra la que presenta mayores cifras de esta según el Minsa [6].

En 2002, el portal del Gobierno del Perú afirmó que el 11 % de la población del Perú padecía ERC [14]. En 2024, el portal web Infobae informó que más de 2,5 millones sufren de esta enfermedad según el Minsa [5]. Esto convierte a los riñones en uno de los órganos más demandados para trasplantes.

El objetivo de este trabajo es encontrar la demanda de atención de pacientes con ERC en el Perú en 2022. Para ello, se emplearán modelos de series de tiempo para predecir dicha demanda, eligiendo el mejor. Además, se ejecutará análisis descriptivo con la data de los pacientes del Fondo Intangible Solidario de Salud.

FUNDAMENTOS

Jilani *et al.* [10] buscan desarrollar un modelo para predecir la atención a los servicios de urgencia, a fin de planificar de manera óptima los recursos. Así, concluyen que el modelo de lógica difusa basado en series de tiempo difusas tuvo una precisión aceptable durante un periodo corto de tiempo a comparación de otros modelos de predicción que se utilizan comúnmente. Por otra parte, Darío y Martínez [7] buscan identificar un buen modelo de pronóstico usando técnicas de suavización y realizar la validación de los supuestos de los residuales. Concluyen que existen diversas técnicas para abordar una serie de tiempo y que la selección debe ser la más simple y comprensible posible, apoyada en el uso de *software* adecuado para realizar múltiples ensayos, así como pedir la opinión de expertos en el tema.

En el trabajo «Time series modelling to forecast prehospital EMS demand for diabetic emergencies» [22], los autores tienen como objetivo modelar las tendencias temporales y proporcionar pronósticos de las asistencias prehospitalarias para emergencias diabéticas, utilizando un análisis de series de tiempo del año 2009 al 2015 con el modelo SARIMA (0, 1, 0, 12) fue el que mejor se ajustó, con un MAPE del 4,2 %. Concluyen que la demanda de servicios médicos de urgencia prehospitalarios para emergencias diabéticas está en aumento, lo cual resulta útil para los proveedores al facilitar la planificación y asignación adecuada de recursos para estos servicios. Además, Jin, Ok y Woong [11] desarrollan y evalúan modelos de series temporales para predecir el número diario de pacientes en un hospital coreano, utilizando datos de 2007 a 2008. Establecieron tres modelos de pronóstico, y el modelo SARIMA resultó ser el más adecuado, con un MAPE de 7,4 %. Concluyen que la incorporación de variables explicativas en un modelo SARIMA multivariante mejora la confiabilidad y precisión de los pronósticos.

En la tesis de Parra [15], se plantea como objetivo encontrar un modelo estructural para predecir la demanda de pacientes en un centro de salud de Guayaquil. La conclusión señala que, para la variable consulta general, el modelo 4, basado en nivel, pendiente, tendencia, estacionalidad, ciclo y componente irregular, es el que mejor se ajusta a los datos. En cambio, para la variable odontología, el modelo 3, basado en la componente estacional, es el que presenta un mejor ajuste. Por su parte, en el estudio de Sen y Chaudhuri [17], el objetivo es aplicar un enfoque de descomposición de series temporales para analizar el comportamiento y las propiedades de la serie. Se proponen seis modelos para predecir indicadores en el sector salud, y se concluye que estos alcanzan un nivel de precisión aceptable, incluso ante la presencia de componentes aleatorios y tendencias abruptas.

Rosa-Jiménez, Montijano, Ília y Zambrana [16] analizan la atención primaria y las urgencias tomando en cuenta las características epidemiológicas de los pacientes que acuden a consultas externas en diversas especialidades. Concluyen que existe una mayor demanda sanitaria por parte de las mujeres, aunque esta varía según la edad, el lugar de procedencia y el origen de la solicitud de asistencia.

Por todo lo anterior, este trabajo tiene como propósito pronosticar la demanda de pacientes con enfermedad renal crónica en los establecimientos de salud estatales del Perú para 2022, mediante modelos de series de tiempo, y realizar un análisis descriptivo de dicha demanda en función de características específicas.

La limitación más importante es que solo se encuentran datos sobre atenciones de cobertura de enfermedad renal de 2022 y el primer semestre de 2023; por ese motivo, se propondrán modelos de series de tiempo en intervalos diarios para tener con una mayor cantidad de datos.

METODOLOGÍA

Tipo y diseño de investigación

La investigación que se llevará a cabo es de tipo cuantitativa [8]. Se trata de un estudio de alcance descriptivo, ya que busca especificar las propiedades y características de un fenómeno, y

exploratorio, dado que examina un tema poco estudiado o no abordado previamente [8]. El objetivo es pronosticar la demanda de atención de pacientes con enfermedad renal crónica en 2022 y realizar un análisis descriptivo de dicha demanda.

El diseño es no experimental, dado que «no se manipulan de forma intencional las variables independientes para observar su efecto en otras variables» [8]. Asimismo, es transversal, puesto que los datos se tomaron en un solo momento en el tiempo, y descriptivo, ya que indaga sobre los niveles de una o más variables en la población [8].

Dado que Hernández, Fernández Baptista (2014) [8] es una referencia destacada en el campo de la investigación, se ha optado por seguir su metodología. No obstante, es importante tener en cuenta que existen diversas clasificaciones y métodos de investigación [18].

Población y muestra

La población de estudio está conformada por 1 064 744 registros de atenciones de cobertura de ERC de 2022 y cuyas variables se detallan en la tabla 1.

Tabla 1
Variables, descripción, tipo de dato y tamaño de atenciones de cobertura de enfermedad renal crónica, 2022

Variable	Descripción	Tipo de dato	Tamaño
DOCUMENTO_ANONIMIZADO	Documento del paciente anonimizado	Alfanumérico	10
PERIODO	Periodo de la atención del paciente oncológico	Numérico	6
RENAES	Código de identificación del establecimiento de salud	Texto	10
IPRESS	Nombre del establecimiento de salud donde se realizó la atención oncológica	Texto	80
REGIÓN	Región donde se encuentra ubicada el establecimiento de salud	Texto	30
DEPARTAMENTO	Departamento donde se encuentra ubicado el establecimiento de salud	Texto	15
PROVINCIA	Provincia donde se encuentra ubicado el establecimiento de salud	Texto	100
DISTRITO	Distrito donde se encuentra ubicado el establecimiento de salud	Texto	100
UBIGEO	Código de ubicación geográfica donde se encuentra ubicado el establecimiento de salud	Alfanumérico	6
CÓDIGO_DIAGNÓSTICO	Código CIE-10 (xxxxxxx) con el cual se identifica al diagnóstico del paciente	Texto	8
DIAGNÓSTICOS	Descripción del diagnóstico CIE-10	Texto	100
GRUPO_DIAGNÓSTICOS	Grupo de diagnóstico al cual pertenece el CIE-10	Texto	100
GRUPO_COBERTURA	El grupo de las categorías de los grupos de diagnósticos	Texto	100
SEXO	Sexo del paciente oncológico	Texto	10
EDAD	Edad del paciente oncológico	Numérico	3
TIPO_SEGURO	Tipo de seguro de cobertura del FISSAL	Texto	30
SERVICIO	Descripción del código de la prestación o atención	Texto	100
FECHA_ATENCIÓN	Fecha en que se brindó la atención	Numérico	8
MONTO_BRUTO	Monto total de la prestación oncológica	Numérico	24
FECHA_CORTE	Día en el que se generó el dataset	Numérico	8

Fuente: Plataforma de Datos Abiertos, Fondo Intangible Solidario de Salud, 2022.

Debido a que se construirán modelos de series de tiempo en intervalos diarios, no se realizará un muestreo. Esto con el objetivo de tener una mayor cantidad de datos, pudiendo decir que el muestreo llevado a cabo fue no probabilístico por conveniencia que, según [4], «el muestreo probabilístico resulta excesivamente costoso y se acude a métodos no probabilísticos, aun siendo conscientes de que no sirven para realizar generalizaciones, pues no se tiene certeza de que la muestra extraída es representativa».

Técnicas de recolección de datos

Se recurrió a una fuente de datos secundaria y tal como lo define Malhotra [12], «las fuentes de datos secundarios externos publicados incluyen agencias gubernamentales federales, estatales y locales» [12]. Los datos fueron tomados de la página web de la Plataforma de Datos Abiertos[1]. Debido a lo anterior, no se requirió de un instrumento de recolección de datos, y la información se recogió en un archivo en formato CSV y que contiene información sobre atenciones de cobertura de enfermedad renal crónica que van del 2 de enero de 2022 al 31 de diciembre de 2022.

Análisis de los datos

El análisis de la información se realizó mediante gráficos de barras, gráficos de barras apiladas, gráficos circulares y tablas, con el fin de observar las frecuencias en las categorías, tanto para las variables numéricas como no numéricas. Además, estas herramientas se emplearon para mostrar los resultados obtenidos de los modelos de series de tiempo, que analizan datos recopilados en intervalos sucesivos para hacer predicciones. También se incluyeron los valores de los hiperparámetros, que se establecen antes del proceso de aprendizaje de un modelo, y las métricas de desempeño, indicadores cuantitativos que evalúan la eficacia de un modelo de predicción.

El *software* utilizado para todo el análisis fue Python en su versión 3.11.5, a través de la interfaz web Jupyter Notebook versión 7.0.6, perteneciente a la distribución Anaconda. Las librerías empleadas fueron pandas, para trabajar con *data frames* en

1 Véase: https://www.datosabiertos.gob.pe/dataset/atenciones-de-cobertura-de-enfermedad-renal-cr%C3%B3nica-2022-fondo-intangible-solidario-de-salud

forma de tablas; matplotlib, para la creación de visualizaciones; numpy, para el manejo de arreglos multidimensionales (matrices); sklearn, para el preprocesamiento, la creación de modelos de aprendizaje automático y el cálculo de métricas de desempeño; y skforecast, para realizar predicciones a uno o varios pasos en el futuro mediante técnicas de pronóstico.

RESULTADOS

Se realizó un análisis descriptivo de la demanda, con base en las variables más relevantes de la tabla 1.

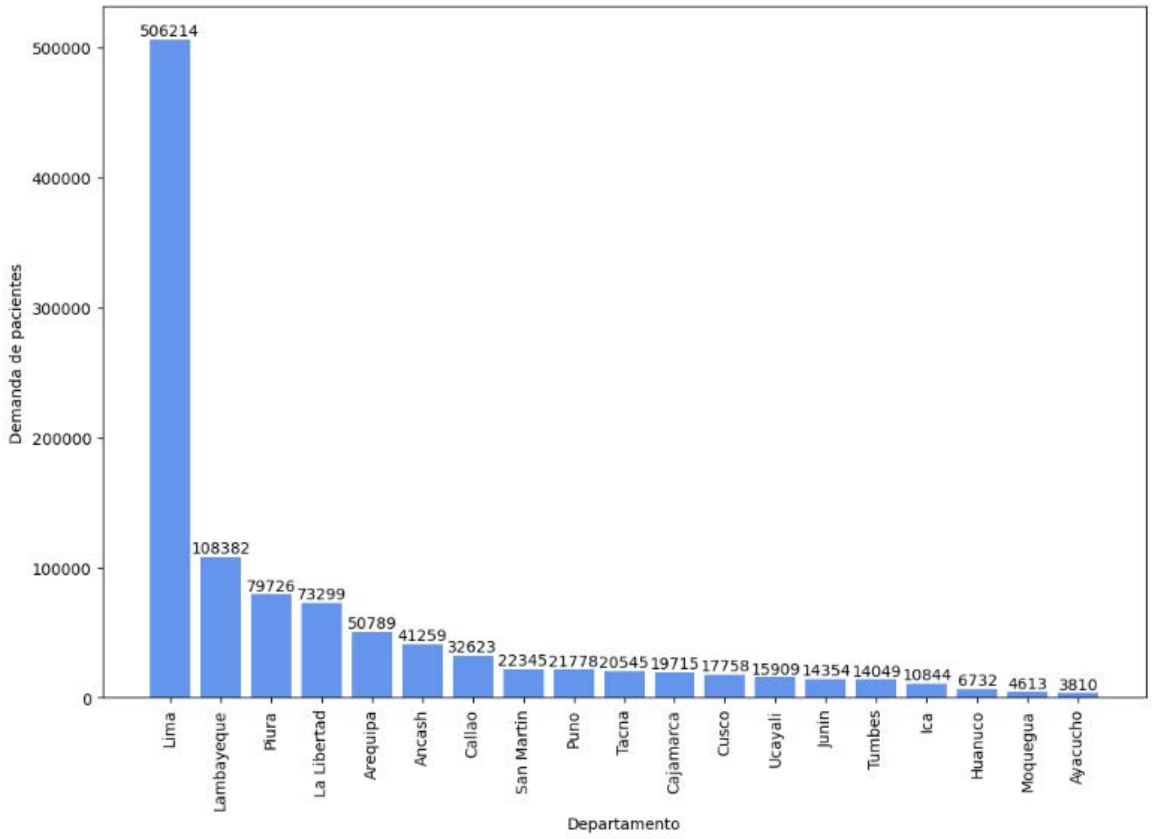


Figura 1. Gráfico de barra de la demanda de pacientes con enfermedad renal crónica por departamento, 2022.

Fuente: Elaboración propia.

En la figura 1, podemos observar que Lima cuenta con la mayor demanda, siendo esta de 506 214; mientras que Ayacucho, con la menor demanda, con un total de 3810.

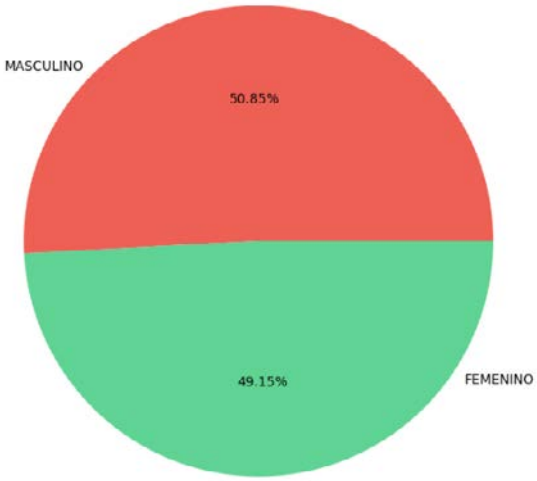


Figura 2. Gráfico circular de la demanda de pacientes con enfermedad renal crónica por sexo, 2022

Fuente: Elaboración propia.

Con respecto a la figura 2, vemos que existe un mayor porcentaje de demanda de pacientes con enfermedad renal crónica del sexo masculino (50,85 %) a diferencia del sexo femenino (49,15%), aunque la diferencia no es mucha.

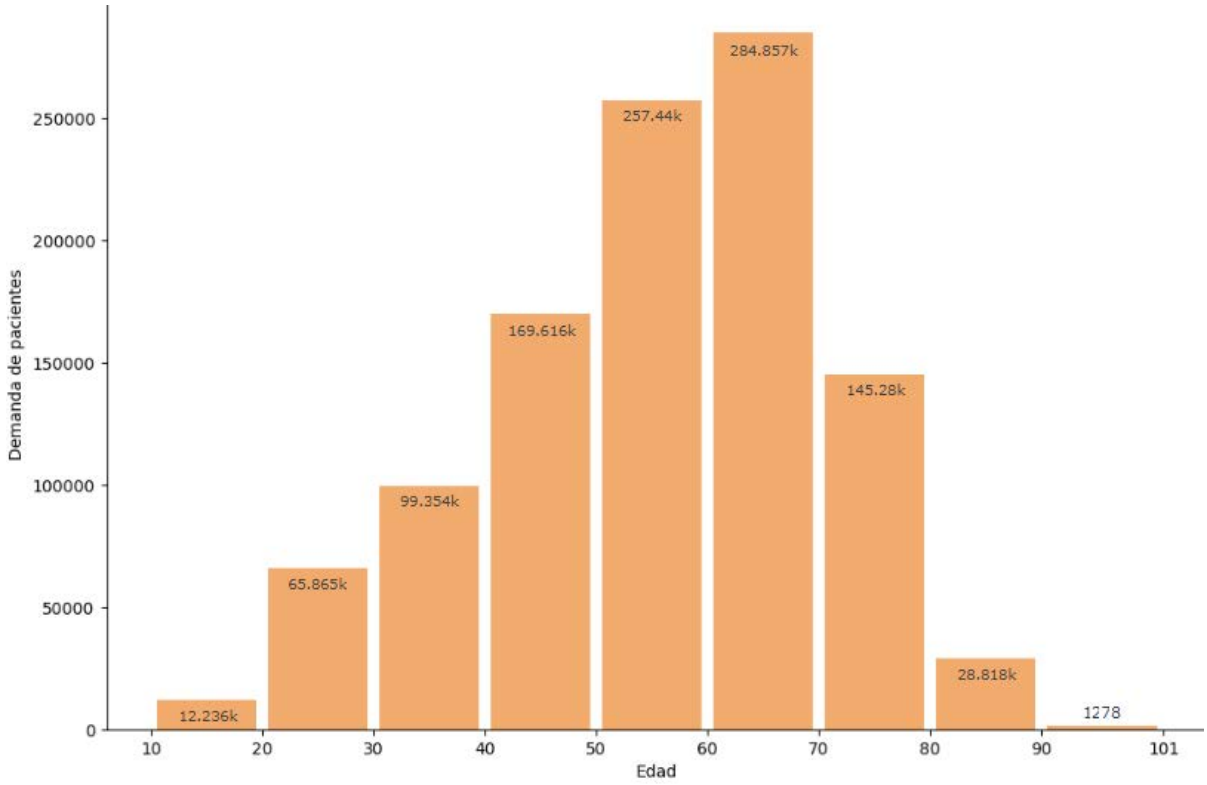


Figura 3. Gráfico de barra de la demanda de pacientes con enfermedad renal crónica por rango de edades, 2022

Fuente: Elaboración propia.

En la figura 3, vemos que la mayor demanda de pacientes con enfermedad renal crónica se encuentra en el rango de edades que va de 50 a 60 años y de 60 a 70 años, sumando un total de 542 297. Por otro lado, la menor demanda se encuentra en el rango de edades de 90 a 101 años, siendo esta de 1278.

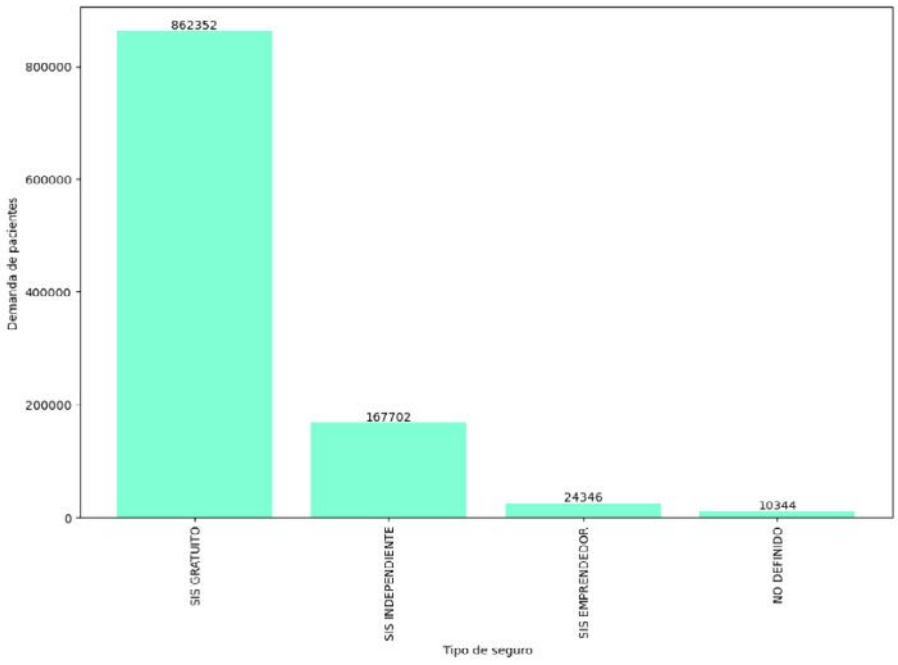


Figura 4. Gráfico de barra de la demanda de pacientes con enfermedad renal crónica por tipo de seguro, 2022

Fuente: Elaboración propia.

En la figura 4, vemos que en el SIS gratuito se encuentra la mayor demanda de pacientes con enfermedad renal crónica, siendo esta de 862 352 y en el SIS emprendedor se encuentra la menor demanda representando 167 702. Además, se encontraron 10 344 registros en los cuales no se especifica el tipo de seguro.

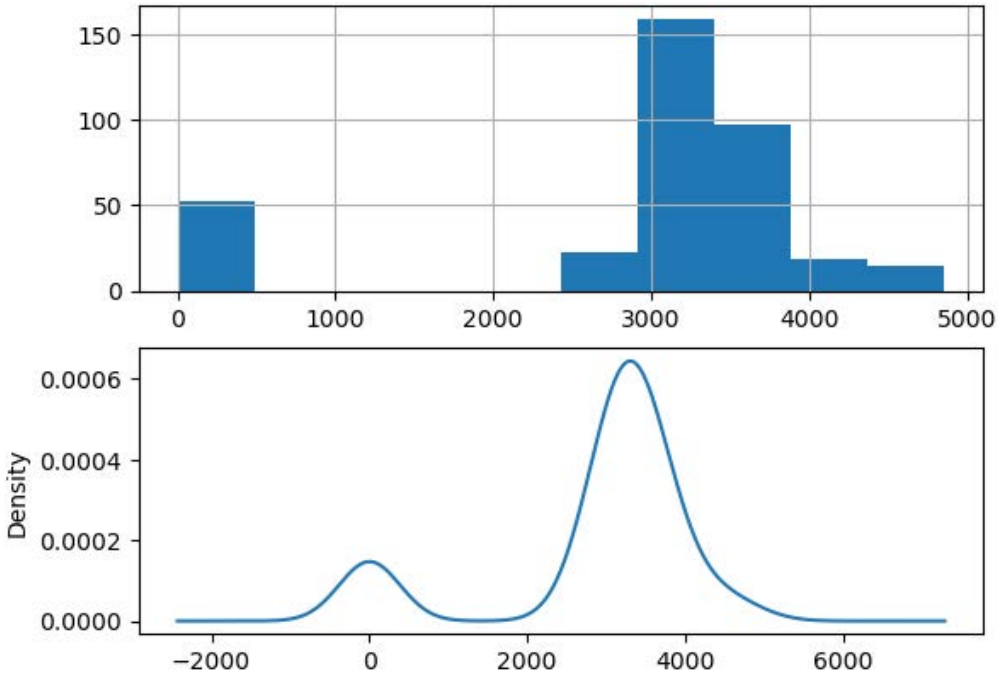


Figura 5. Histograma y gráfico de densidad de la demanda por atención de pacientes con enfermedad renal crónica, 2022

Fuente: Elaboración propia.

En la figura 5, observamos que, de acuerdo con el histograma y gráfico de densidad, el número de atenciones parece no seguir una distribución normal, sino, más bien, una distribución mixta normal.

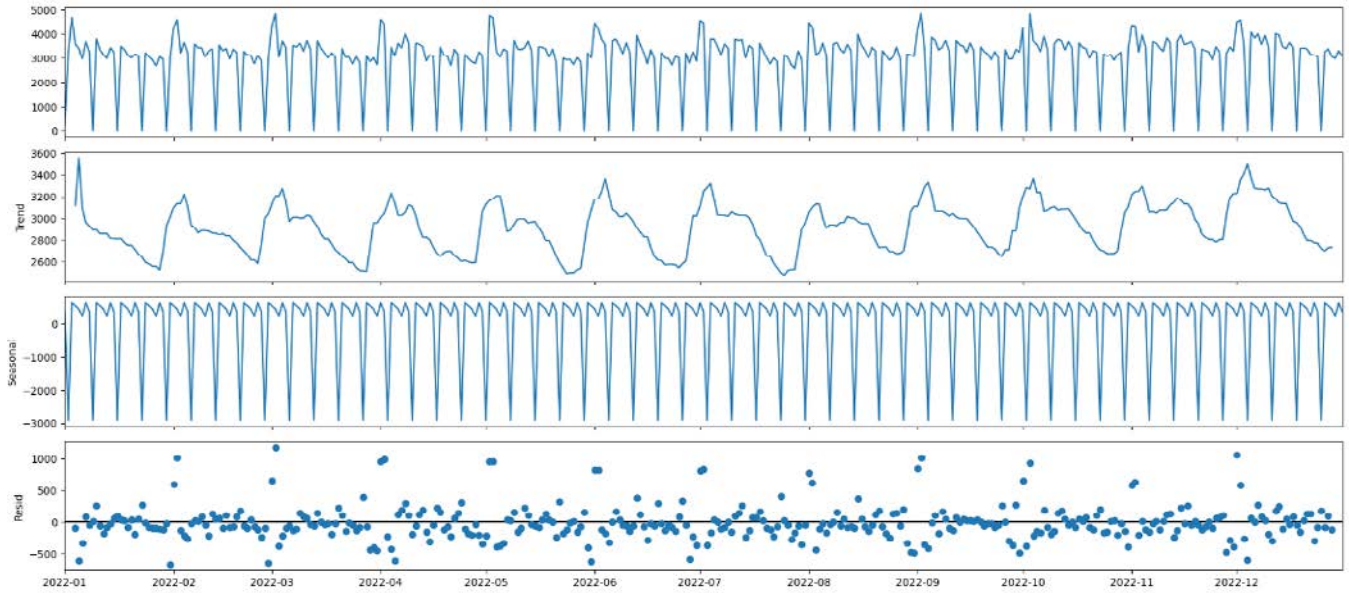


Figura 6. Gráfico de tendencia, estacionalidad y residuales de la demanda por atención de pacientes con enfermedad renal crónica, 2022

Fuente: Elaboración propia.

De acuerdo con la figura 6, se observa una ligera tendencia a crecer, manteniendo una estacionalidad casi constante y cuyos residuales parecen seguir un comportamiento aleatorio.

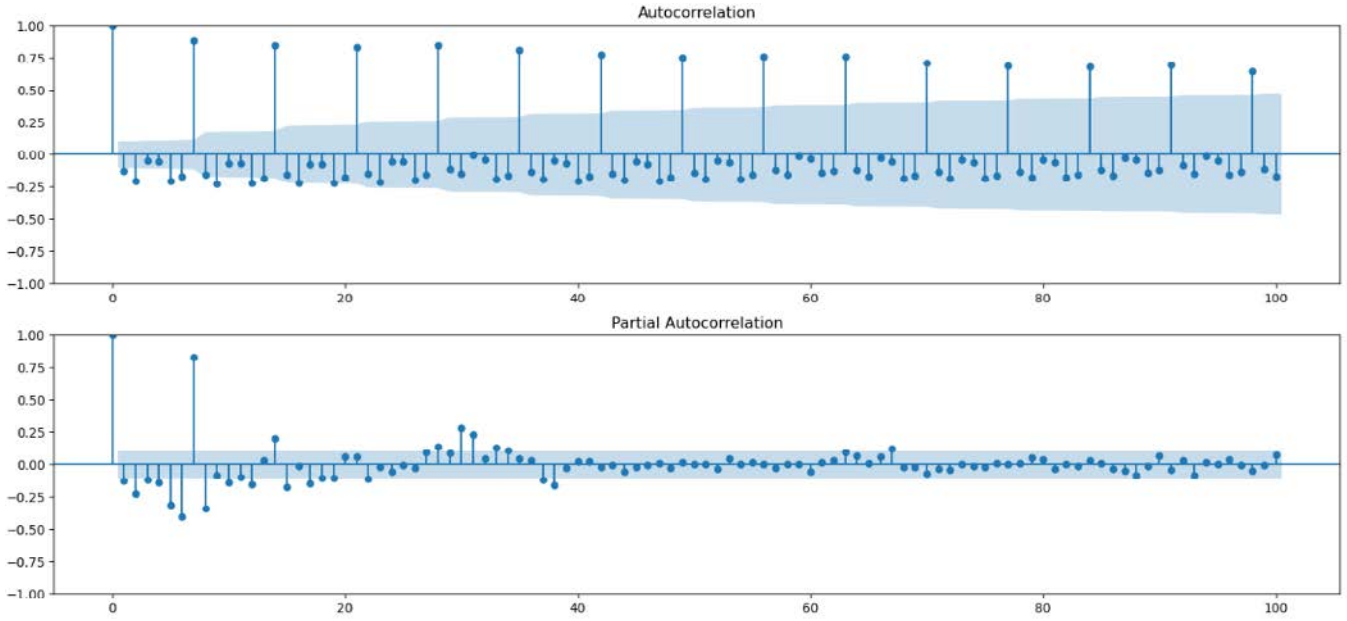


Figura 7. Gráfico de autocorrelación y autocorrelación parcial de la demanda por atención de pacientes con enfermedad renal crónica, 2022

Fuente: Elaboración propia.

En la figura 7, vemos que existe correlación con retardos de 7. Esto debido al hecho de que no se llevan a cabo atenciones por enfermedad renal crónica los domingos.

Tabla 2
Resultados de la prueba de Dickey-Fuller

Valor del estadístico	p-valor	Número de lags usados
-6,47	1,37e-08	17

Fuente: Elaboración propia.

De acuerdo con la tabla 2, con un p-valor menor a un nivel de significancia de 0,05, debemos rechazar la hipótesis nula de la prueba de Dickey-Fuller la cual afirma que los datos no son estacionarios. Así, lo recomendable sería proceder sin usar diferencias.

De acuerdo con la figura 8, notamos que existen demandas de atenciones que son iguales a cero, lo cual se debe a que dichas atenciones se realizaron un domingo, donde, generalmente, no se presta ese servicio en los locales de salud.

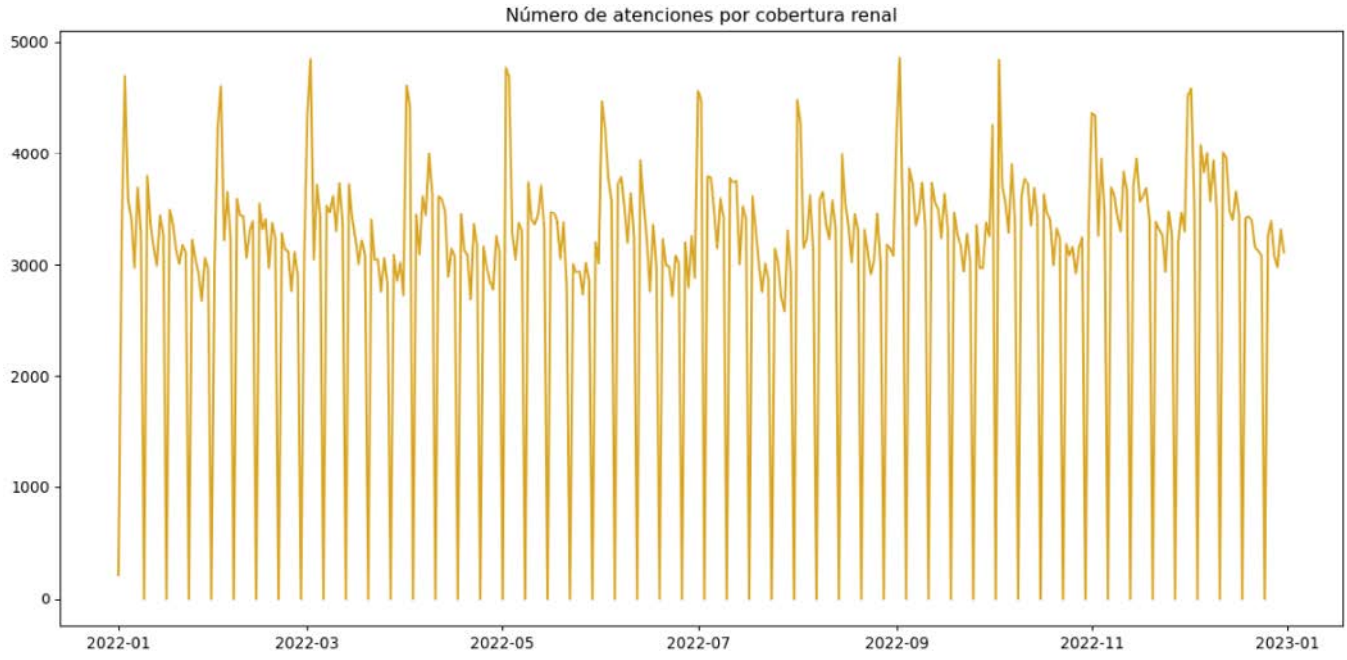


Figura 8. Gráfico de líneas de la demanda por atención de pacientes con enfermedad renal crónica, 2022

Fuente: Elaboración propia.

Tabla 3
Valores atípicos para la demanda por atención de pacientes con enfermedad renal crónica, 2022

Fecha_atención	Número_atenciones
2022-01-01	208
2022-01-02	3256
2022-03-27	1
2022-09-11	1

Fuente: Elaboración propia.

De acuerdo con la tabla 3, en la fecha de atención del 2022-01-01 se tiene una demanda de atención de 208 a pesar de ser feriado y en la fecha de atención del 2022-01-02 se tiene una demanda de 3256 a pesar de ser domingo. Además, notamos que las demandas de atenciones con fechas de atención del 2022-01-27 y 2022-09-11 con valores 1 y 1, respectivamente, son diferentes de cero a pesar de ser domingo, aunque son pequeños a comparación de la mayoría.

Para entrenar nuestro modelo, se tomará los 30 últimos días para los datos de prueba y el resto para el entrenamiento. Se buscará predecir el número de atenciones para estos 30 días, por medio de predicciones *multistep* recursivas, en la cual cada nueva predicción utiliza la predicción anterior [3].

Tabla 4
Valores para los parámetros para el modelo *forecasting* autorregresivo recursivo

Regresor	Número de lags usados	Rango de entrenamiento
Random Forest	7	2022-01-01 al 2022-12-01

Fuente: Elaboración propia.

En la tabla 4, podemos observar los valores para los parámetros de nuestro modelo en la cual el regresor es un modelo *Random Forest*, el número de lags será de 7, que es lo que comúnmente se repite cada semana y tomando datos de entrenamiento desde el 1 de enero hasta el 1 de diciembre de 2022.

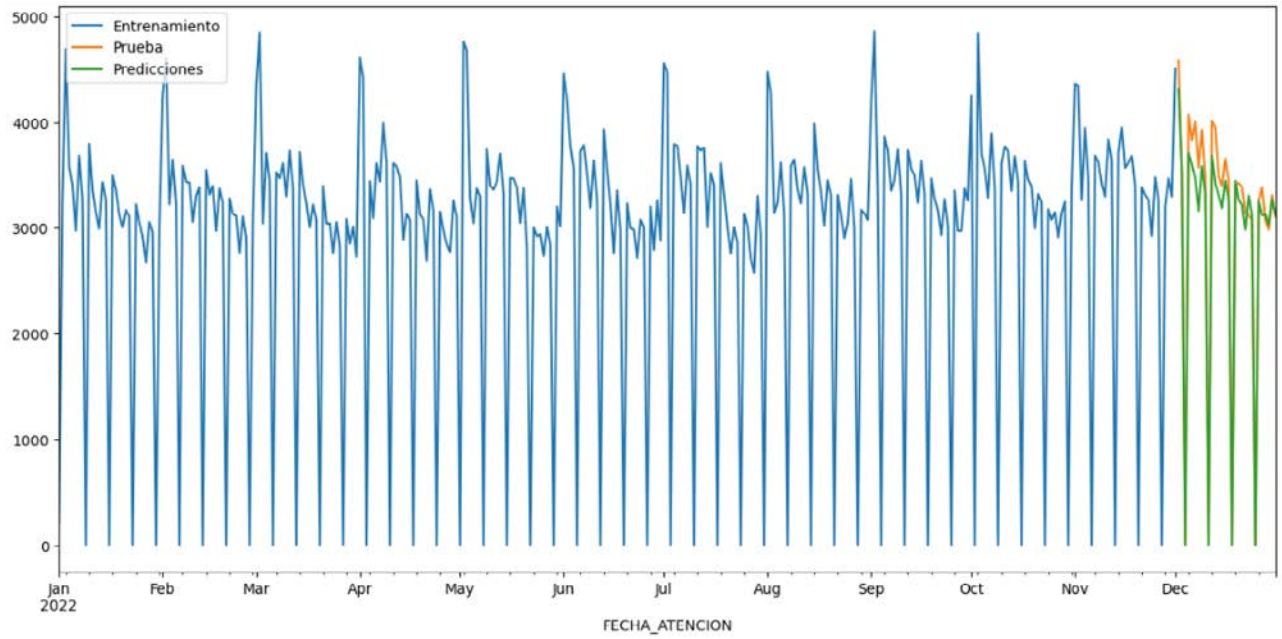


Figura 9. Gráfico de líneas para el entrenamiento, prueba y predicciones de la demanda por atención de pacientes con enfermedad renal crónica, 2022

Fuente: Elaboración propia.

De la figura 9, se evidencia que las predicciones siguen un comportamiento similar a los datos de prueba, lo cual nos da un indicio, de que el modelo entrenado es bueno para predecir el número de atención de pacientes.

Tabla 5
Valores de las métricas de desempeño para el modelo *forecasting* autorregresivo recursivo

Métrica	Valor
Error cuadrático medio	52 728,29
Raíz del error cuadrático medio	229,62
Error absoluto medio	171,36
Coeficiente de determinación	0,9662

Fuente: Elaboración propia.

En la tabla 5, observamos los valores de las métricas más importantes para nuestro modelo *forecasting* autorregresivo recursivo de donde podemos ver que posee un coeficiente de determinación de 0,9662, lo cual nos indica que es un buen modelo para predecir.

Tabla 6
Valores de los parámetros y el error cuadrático medio para varios modelos *forecasting* autorregresivo recursivo

Parámetros			
Lags	**Máxima profundidad**	**Número de estimadores**	**Error cuadrático medio**
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	7	100	149 152,4
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	5	100	149 490,2
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	7	500	149 842,7
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	5	500	149 898,7
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	5	300	151 203,0
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	3	500	153 169,6
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	3	300	153 238,0
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	3	100	153 331,4
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	7	300	154 554,5
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	10	100	154 832,4
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	10	500	159 258,1
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	10	300	162 457,3
[1, 2, 3, 4, 5]	5	500	833 534,3
[1, 2, 3, 4, 5]	5	300	1 118 611
[1, 2, 3, 4, 5]	5	100	1 170 591
[1, 2, 3, 4, 5]	7	100	1 366 903
[1, 2, 3, 4, 5]	10	300	1 424 055
[1, 2, 3, 4, 5]	7	500	1 463 126
[1, 2, 3, 4, 5]	10	100	1 467 230
[1, 2, 3, 4, 5]	7	300	1 523 321
[1, 2, 3, 4, 5]	10	500	1 739 128
[1, 2, 3, 4, 5]	3	300	2 091 496
[1, 2, 3, 4, 5]	3	100	2 176 777
[1, 2, 3, 4, 5]	3	500	2 206 203

Fuente: Elaboración propia.

De acuerdo con los resultados de la tabla 6, al realizar el ajuste de hiperparámetros, el mejor modelo es el generado mediante 12 lags, con una máxima profundidad de 7 y un número de estimadores de 100, siendo el error cuadrático medio de 149 152,4.

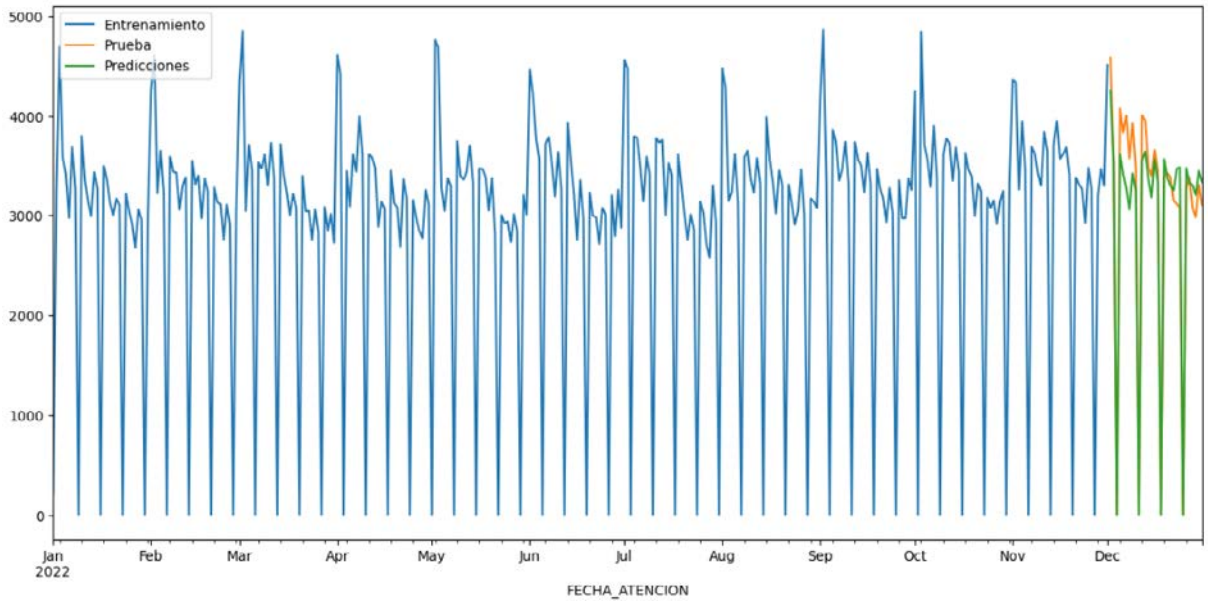


Figura 10. Gráfico de líneas para el entrenamiento, prueba y predicciones con ajuste de hiperparámetros de la demanda por atención de pacientes con enfermedad renal crónica, 2022

Fuente: Elaboración propia.

De la figura 10, se observa que las predicciones con ajuste de hiperparámetros siguen un comportamiento similar a los datos de prueba. Esto nos da un indicio de que el modelo entrenado es bueno para predecir el número de atención de pacientes.

Tabla 7
Valores de las métricas de desempeño para el modelo *forecasting* autorregresivo recursivo con ajuste de hiperparámetros

Métrica	Valor
Error cuadrático medio	84 123,96
Raíz del error cuadrático medio	290,04
Error absoluto medio	224,08
Coeficiente de determinación	0,9461

Fuente: Elaboración propia.

En la tabla 7, se presentan las métricas más relevantes para nuestro modelo *forecasting* autorregresivo recursivo con ajuste de hiperparámetros. Observamos que el coeficiente de determinación es de 0,9461, lo que indica que el modelo sigue siendo eficaz para realizar predicciones, aunque con menor sobreajuste, lo cual es favorable para la generalización de los resultados.

En la figura 11, observamos que, en el gráfico de valores predichos vs. valores reales, los puntos están distribuidos aproximadamente alrededor de la recta identidad, lo que sugiere que se cumple la suposición de linealidad. Por otro lado, en el gráfico de residuos vs. órdenes, los residuos parecen mostrar un comportamiento aleatorio, lo que indicaría que se cumple el supuesto de independencia. El gráfico de distribución de los residuos propone que estos siguen, aproximadamente, una distribución normal. Sin embargo, en el gráfico de residuos vs. valores predichos, notamos que no todos los puntos se encuentran alrededor de cero, lo que indica que no se cumple la homocedasticidad [19].

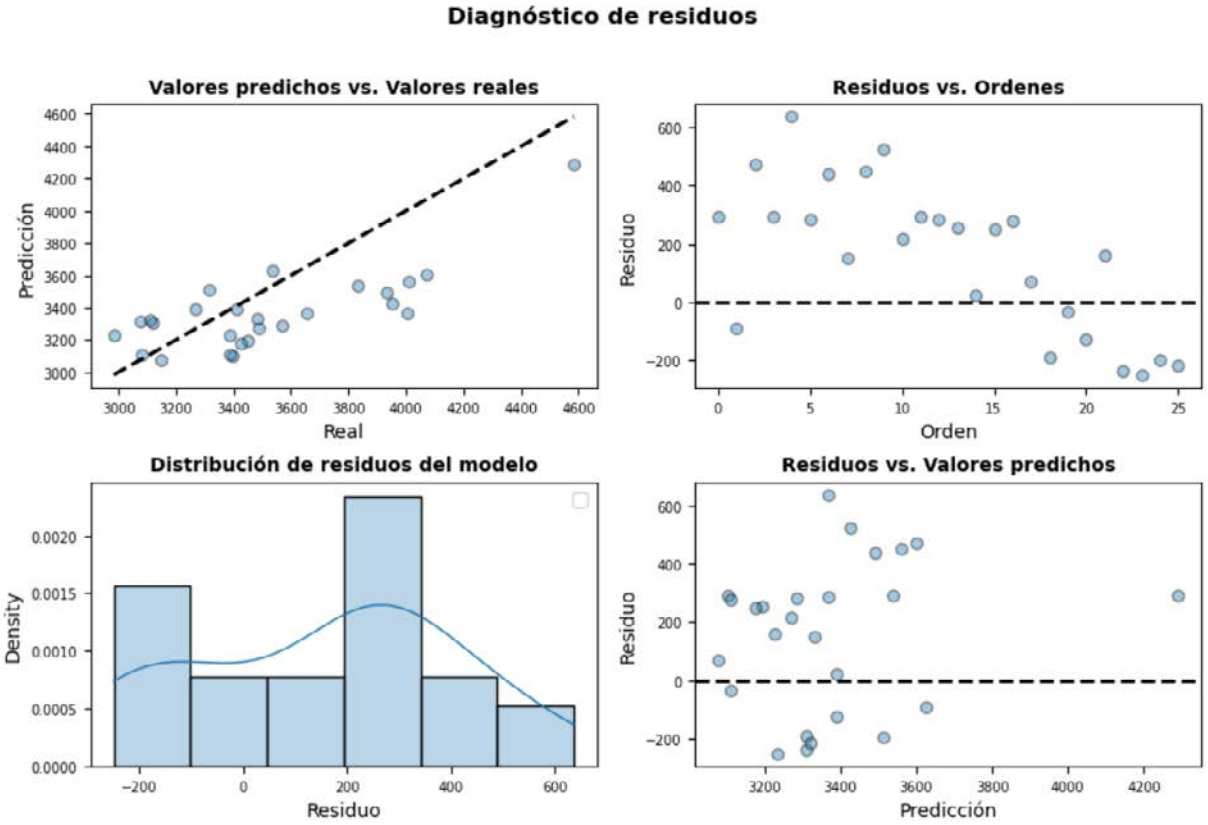


Figura 11. Diagnóstico de residuos para el modelo *forecasting* autorregresivo recursivo con ajuste de hiperparámetros

Fuente: Elaboración propia.

Tabla 8
Resultados de la prueba de Shapiro-Wilk

Valor del estadístico	p-valor
0,94	0,1399

Fuente: Elaboración propia.

De acuerdo con la tabla 8, con un p-valor menor a un nivel de significancia de 0,05, debemos aceptar la hipótesis nula de la prueba de Shapiro-Wilk. Esta afirma que los residuos cumplen el supuesto de normalidad [1].

Autocorrelación positiva	Zona de indecisión	Zona de no autocorrelación	Zona de indecisión	Autocorrelación negativa
0 d_L d_U 2 4 – d_U 4 – d_L 4				

Figura 11. Rango de valores del estadístico Durbin-Watson para la prueba de no autocorrelación

Fuente: Elaboración propia.

En la figura 11, vemos el rango de valores del estadístico Durbin-Watson que utilizaremos para la prueba de no autocorrelación. Si buscamos en una tabla de valores críticos del estadístico Durbin-Watson [13], encontramos que, para dos términos que incluyen el intercepto, los valores críticos son $d_L = 1,35$ y $d_U = 1,49$ con un nivel de significancia de 0,05.

Tabla 9
Resultados de la prueba de Durbin-Watson

Valor del estadístico
0,67

Fuente: Elaboración propia.

En la tabla 9, con un nivel de significancia de 0,05 y con un valor estadístico de d = 0,67, que cumple d < 1.35, se evidencia que los residuos están autocorrelacionados positivamente, con lo cual no cumplen el supuesto de independencia.

Tabla 10
Resultados de la prueba de Bartlett

Valor del estadístico	p-valor
4,94	0,0263

Fuente: Elaboración propia.

Con base en la tabla 10, con un p-valor menor a un nivel de significancia de 0,05, debemos rechazar la hipótesis nula de la prueba de Bartlett la cual afirma que los residuos tienen varianza constante. No obstante, si tomamos un nivel de significancia de 0,01, se aceptará la hipótesis nula de la prueba de Bartlett, con lo cual los residuos cumplen el supuesto de homocedasticidad [2].

CONCLUSIONES

Lima es el departamento con mayor demanda a nivel nacional en comparación con los demás. Ayacucho es el que presenta la menor demanda. Además, la demanda entre hombres y mujeres está equilibrada.

Con relación a la demanda por edades, es mayor en adultos mayores y menor en personas que superan los 90 años. En el seguro gratuito se registra la mayor demanda, mientras que en el seguro emprendedor es la menor. También se encontró una categoría con valores nulos, al no estar definido el tipo de seguro.

El análisis de la demanda a través de una serie de tiempo diaria mostró que los datos siguen una distribución normal mixta, con una tendencia ligeramente creciente, estacionalidad casi constante y residuales de comportamiento aleatorio. Los gráficos de correlación y autocorrelación revelaron una correlación con retardos o lags de 7, dado que en la mayoría de los centros de atención no se brinda servicio los domingos.

La prueba de Dickey-Fuller indicó que los datos no son estacionarios, por lo que no es recomendable proceder con diferencias. Se detectaron algunos valores atípicos en cuatro fechas específicas, correspondientes a feriados y domingos, que no fueron omitidos, ya que no presentaban valores muy grandes.

Se entrenó un modelo *forecasting* autorregresivo recursivo con 7 retardos o lags y haciendo uso de un regresor Random Forest, en donde se observó gráficamente que las predicciones tienen un comportamiento similar a los datos de prueba con un valor de coeficiente de determinación bastante bueno (96,62 %).

Debido a un posible sobreajuste en el modelo, se llevó a cabo un ajuste de hiperparámetros tomando como parámetros, la cantidad de lags, la máxima profundidad y el número de estimadores. En el modelo ganador, las predicciones también tienen un comportamiento similar a los datos de prueba, aunque esta vez con un valor de coeficiente de determinación menor, pero que sigue siendo bueno (94,61 %).

Al realizar un análisis de los supuestos del modelo, se observó de manera gráfica que existe una relación lineal entre los valores reales y predichos; la prueba de Shapiro-Wilk demostró que los residuos siguen una distribución normal; con respecto a la independencia, la prueba de Durbin-Watson dio como resultado que existe cierto nivel de autocorrelación; y en el caso de la homocedasticidad de los residuos, la prueba de Bartlett mostró que esto no se cumple, a menos que se utilice un nivel de significancia mayor.

Se concluye que el modelo propuesto tiene un buen desempeño, a pesar de no cumplir todos los supuestos del modelo *forecasting* autorregresivo recursivo, ya que solo falla en la independencia. Es útil para predecir la demanda de atención de pacientes con ERC en los meses o años posteriores a 2022.

Por último, sobre las limitaciones encontradas en el estudio, no se contó con un mayor rango de años para obtener un modelo más preciso. Se podría optar por realizar futuras investigaciones con respecto a otras demandas de enfermedades crónicas como la diabetes y tipos de cáncer. Además, de acuerdo con lo mencionado por Villani *et al*. [20], este trabajo puede ser útil para que las entidades de salud gubernamentales ejecuten una mejor planificación y asignación de recursos en los servicios médicos hospitalarios.

REFERENCIAS

[1] Amat, J. (2021a). *Ciencia de datos con Python: Análisis de normalidad con Python*. Ciencia de Datos. https://cienciadedatos.net/documentos/pystats06-analisis-normalidad-python

[2] Amat, J. (2021b). *Ciencia de datos con Python: Análisis de homocedasticidad y heterocedasticidad con Python*. Ciencia de Datos. https://cienciadedatos.net/documentos/pystats07-test-homocedasticidad-heterocedasticidad-python

[3] Amat, J. & Escobar, J. (2023). *Skforecast: Forecasting series temporales con Python y Scikit-learn*. Ciencia de Datos. https://cienciadedatos.net/documentos/py27-forecasting-series-temporales-python-scikitlearn.html

[4] Arias-Gómez, J., Villasís-Keever, M. & Miranda, M. (2016). El protocolo de investigación III: La población de estudio. *Alergia México*, *63*(4), 201-206.

[5] Campó, S. (14 de marzo de 2024). En Perú más de 2,5 millones sufren de enfermedad renal crónica, según el Minsa: ¿Cómo cuidar la salud de los riñones? *Infobae*. https://www.infobae.com/peru/2024/03/14/en-peru-

mas-de-25-millones-sufren-de-enfermedad-renal-cronica-segun-el-minsa-como-cuidar-la-salud-de-los-rinones/

[6] Carrillo-Larco, R. & Berbané-Ortiz, A. (2018). Mortalidad por enfermedad renal crónica en el Perú: Tendencias nacionales 2003-2015. *Revista Peruana de Medicina Experimental y Salud Pública*, *35*(3), 409-415.

[7] Darío, L. & Martínez, S. (2007). Una metodología de series de tiempo para el área de la salud: Caso práctico. *Facultad Nacional de Salud Pública*, *25*(2), 117-122.

[8] Hernández, R., Fernández, C. & Baptista, M. (2014). *Metodología de la investigación* (6.ª ed.). McGraw-Hill Education.

[9] Herrera-Añazco, P., Pacheco-Mendoza, J. & Taype-Rondán, A. (2016). La enfermedad renal crónica en el Perú: Una revisión narrativa de los artículos científicos publicados. *Acta Médica Peruana*, *33*(2), 130-137.

[10] Jilani, T. *et al.* (2019). Short and long term predictions of hospital emergency department attendances. *International Journal of Medical Informatics*, *129*, 167-174.

[11] Jin Kam, H., Ok, S. & Woong, P. (2010). Prediction of daily patient numbers for a regional emergency medical center using time series analysis. *Healthcare Informatics Research*, *16*(3), 158-165.

[12] Malhotra, N. (2008). *Investigación de mercados* (5.ª ed.). Pearson Educación.

[13] Minitab. (2024). Comprobar si existe autocorrelación usando el estadístico de Durbin-Watson. *Soporte de Minitab*. https://support.minitab.com/es-mx/minitab/help-and-how-to/statistical-modeling/regression/supporting-topics/model-assumptions/test-for-autocorrelation-by-using-the-durbin-watson-statistic/#fntarg_1

[14] Ministerio de Salud del Perú [Minsa]. (10 de marzo de 2022). Día Mundial del Riñón: El 11 % de la población del Perú padece una enfermedad renal crónica. *Gobierno del Perú*. https://www.gob.pe/institucion/minsa/noticias/589662-dia-mundial-del-rinon-el-11-de-la-poblacion-del-peru-padece-una-enfermedad-renal-cronica

[15] Parra, R. (2012). *Un modelo estructural de series de tiempo para la predicción de la demanda de atención médica en el sistema municipal de salud* [Tesis de maestría, Escuela Superior Politécnica del Litoral]. Repositorio Institucional ESPOL.

[16] Rosa-Jiménez, F., Montijano, A., Ília, C. & Zambrana, J. (2005). ¿Solicitan las mujeres más consultas al área médica? *Anales de Medicina Interna*, *22*(11), 515-519.

[17] Sen, J. & Chaudhuri, T. (2017). A time series analysis-based forecasting framework for the Indian healthcare sector. *Journal of Insurance and Financial Management*, *2*(2), 15-34.

[18] Tam, J., Vega, G. & Oliveros, R. (2008). Tipos, métodos y estrategias de investigación científica. *Pensamiento y Acción*, *14*(1), 145-154.

[19] Vásquez, S., Benavides, T. & Ruiz, S. (2023). *Informe series de tiempo*. RPubs. https://rpubs.com/sararuiz/seriesdetiempo

[20] Villani, M. *et al.* (2017). Time series modelling to forecast emergency department presentations. *BMC Health Services Research*, *17*(1), 1-9.

[21] Zanabria-Calderón, J. (2022). Brecha oferta/demanda de prestaciones en el control de la enfermedad renal crónica en EsSalud. *Cátedra Villarreal*, 10(2), 86-97.

ACERCA DEL AUTOR

José Luis Espinoza Melgarejo

Magíster en Docencia Universitaria e Investigación Pedagógica de la Universidad San Pedro de Chimbote (USP), Perú, y licenciado en Matemática de la Universidad Nacional de Ingeniería (UNI), Perú. Docente de educación superior con más de 10 años de experiencia laborando en instituciones como IDAT, Tecsup, la Universidad Privada del Norte (UPN), la Universidad Nacional Mayor de San Marcos (UNMSM) y la Universidad Tecnológica del Perú (UTP). Ha dictado diversos cursos en ciencia de datos, matemática, estadística, finanzas entre otros. Especialista en estadística y ciencia de datos que cuenta con estudios de posgrado en la UNMSM y, actualmente, es estudiante de doctorado en Estadística Matemática de la Universidad Nacional del Santa (UNS), Perú.

jespinozame@tecsup.edu.pe

josespijoin@yahoo.com

Recibido: 21-04-24
Revisado: 22-07-24
Aceptado: 07-08-24