# UNIVERSIDAD NACIONAL AGRARIA

# LA MOLINA

## FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

# "PREDICCIÓN DE VENTAS DE DEPARTAMENTOS EN EL DISTRITO DE MIRAFLORES DE UNA EMPRESA INMOBILIARIA DE LIMA UTILIZANDO EL MODELO DE ENSAMBLE POR MEDIAS"

**TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR TÍTULO DE INGENIERA EN ESTADÍSTICA INFORMÁTICA**

**GABRIELA PILAR CRUZ PAREDES**

**LIMA – PERÚ**

**2024**

---

# TSP_CRUZ

INFORME DE ORIGINALIDAD

| 19% | 18% | 2% | 10% |
|---|---|---|---|
| INDICE DE SIMILITUD | FUENTES DE INTERNET | PUBLICACIONES | TRABAJOS DEL ESTUDIANTE |

ENCONTRAR COINCIDENCIAS CON TODAS LAS FUENTES (SOLO SE IMPRIMIRÁ LA FUENTE SELECCIONADA)

2%

★ addi.ehu.es

Fuente de Internet

| | | | |
|---|---|---|---|
| Excluir citas | Apagado | Excluir coincidencias | < 15 words |
| Excluir bibliografía | Activo | | |

# UNIVERSIDAD NACIONAL AGRARIA

# LA MOLINA

## FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

# "PREDICCIÓN DE VENTAS DE DEPARTAMENTOS EN EL DISTRITO DE MIRAFLORES DE UNA EMPRESA INMOBILIARIA DE LIMA UTILIZANDO EL MODELO DE ENSAMBLE POR MEDIAS"

**PRESENTADO POR:**
**GABRIELA PILAR CRUZ PAREDES**

**TRABAJO DE SUFICIENCIA PROFESIONAL PARA OPTAR EL TÍTULO DE INGENIERA EN ESTADÍSTICA INFORMÁTICA**

SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO

| | |
|---|---|
| Dr. Ivan Dennys Soto Rodriguez<br>PRESIDENTE | Dr. Fernando Rene Rosas Villena<br>ASESOR |
| Dr. Rino Nicanor Sotomayor Ruiz<br>MIEMBRO | Dr. Raphael Felix Valencia Chacón<br>MIEMBRO |

Lima – Perú
2024

## Dedicatoria

A mis padres Julia y Javier, y mis hermanos Katia y Enrique,

que me brindaron su apoyo moral

y motivacional en este tiempo.

## Agradecimiento

A los profesores de la UNALM que durante todo tiempo inspiraron a querer la carrera, y a mis amigos que me enseñaron a perseverar y persistir en el camino de la vida.

# ÍNDICE GENERAL

# ÍNDICE DE TABLAS

# ÍNDICE DE FIGURAS

# RESUMEN

El mercado inmobiliario es un indicador de desarrollo importante en la economía de los países, cuyos principales factores que afectan a este mercado son los económicos y financieros, pero también los políticos pesan mucho, además de las condiciones propias de cada mercado. En el Perú, la situación del mercado inmobiliario fue afectado por la pandemia del COVID19 debido a la paralización de obras de construcción atrasando la entrega de los inmuebles, firmas de minutas con clientes, desembolsos crediticios y otros factores más, razones por la cual empresas del rubro inmobiliario empezaron a cambiar sus estrategias para llegar al cliente final. A pesar de la coyuntura, el mercado inmobiliario ha sido uno de los mercados resilientes, pues la necesidad de vivienda sigue siendo de importancia para las personas, sea para uso propio o para generar otros ingresos. En el presente trabajo se utilizó el modelo de ensamble de medias para estimar las futuras ventas de inmuebles dentro del distrito de Miraflores, el cual primero se usaron el Modelo ARIMA, Modelo de regresión lineal múltiple y por último la combinación de resultados de los modelos por la media simple. Con el método de ensamble por medias se obtuvo un valor de predicción del 85% para las ventas estimadas de departamentos en el distrito de Miraflores, el cual permitirá tomar decisiones en la empresa como continuar invirtiendo en mejorar sus sistemas de marketing, canales de ventas, inversión en terrenos, etc.

**Palabras clave:** mercado inmobiliario, ensamblaje por medias, modelo ARIMA, regresión múltiple, predicción.

# ABSTRACT

The real estate market is an important development indicator in the economy of countries, whose main factors that affect this market are economic and financial, but political factors also weigh heavily, in addition to the conditions of each market. In Peru, the situation of the real estate market was affected by the COVID19 pandemic due to the paralysis of construction works, delaying the delivery of properties, signing of minutes with clients, credit disbursements and other factors, reasons why companies in the real estate sector began to change their strategies to reach the end customer. Despite the situation, the real estate market has been one of the resilient markets, since the need for housing continues to be important for people, whether for their own use or to generate other income. In this work, the ensemble model of means was used to estimate future real estate sales within the Miraflores district, which first used the ARIMA Model, Multiple Linear Regression Model and finally the combination of results from the models by the simple mean. With the assembly method by means, a prediction value of 85% was obtained for the estimated sales of apartments in the Miraflores district, which will allow the company to make decisions such as continuing to invest in improving its marketing systems, sales channels, investment in land, etc.



**Keywords:** real estate market, average ensemble, ARIMA model, multiple regression, prediction.

# I. INTRODUCCIÓN

## 1.1. Problemática

El presente Trabajo de Suficiencia Profesional, fue la predicción de ventas de departamentos en el distrito de Miraflores de una empresa inmobiliaria de Lima utilizando el modelo de ensamble por medias. El estudio se realizó por encargo del Comité de Análisis del Mercado a la Gerencia Comercial, Área de Productos Analíticos y Ciencia de Datos. En dicha área ocupé el cargo de Analista de Productos Analíticos y Ciencia de Datos.

Real State Market de México (2017), portal de investigación del mercado inmobiliario, señala que el mercado inmobiliario es un indicador de desarrollo importante en la economía de los países. Los principales factores que afectan a este mercado son los económicos y financieros, pero también los políticos pesan mucho, además de las condiciones propias de cada mercado.

Durante los últimos años, ha mostrado una tendencia a la baja, después de varios años de fuertes subidas de precios. En EEUU, el precio de la vivienda ha caído un 1.3% mensual en 2022, el doble de lo previsto por los analistas. En otras economías como Suecia, el descenso es incluso mayor, mientras que en las economías en las que aún no se ha materializado una corrección clara hay indicios que apuntan en esa dirección.

El diario Economista de España (2022) indica que una recesión económica puede reducir la demanda de vivienda a través de una mayor tasa de interés. Si las familias o empresas ven reducidos sus ingresos o si prevén un futuro incierto, las decisiones de inversión importantes podrían posponerse, afectando a la demanda y los precios de los inmuebles.

Según reporte de la cadena BBC (2020), mercados como Turquía, Filipinas o Alemania el precio de las casas se disparó en el segundo trimestre de este año, a

pesar de los profundos efectos económicos provocados por la pandemia de COVID19, según un análisis elaborado por la firma Global Property Guide. Como en toda recesión hay ganadores y perdedores, quienes han mantenido su trabajo y contaban con ahorros, están aprovechando que las tasas de interés de los créditos hipotecarios han llegado a niveles históricamente bajos, y esa caída en la tasa de interés ha sido el principal combustible que impulsó la compra de inmuebles.

En el Perú, la situación del mercado inmobiliario fue afectado por la pandemia del COVID19 debido a la paralización de obras de construcción atrasando la entrega de los inmuebles, firmas de minutas con clientes, desembolsos crediticios y otros factores más, razones por la cual empresas del rubro inmobiliario empezaron a cambiar sus estrategias para llegar al cliente final. A pesar de la coyuntura, el mercado inmobiliario ha sido uno de los mercados resilientes, pues la necesidad de vivienda sigue siendo de importancia para las personas, sea para uso propio o para generar otros ingresos.

Según declaraciones del director ejecutivo de CAPECO (Cámara Peruana de Construcción), estimaron un crecimiento del negocio inmobiliario en el Perú sobre el 7% al cierre del 2021. Este crecimiento como consecuencia de un conjunto de factores que cimentaron las bases para que el mercado inmobiliario del 2021 cierre en estas cifras positivas, como: la demanda de unidades familiares, (solo en Lima, ronda los 250.000, contra una oferta que apenas oscila entre las 24.000 y 25.000 unidades), la liberación de CTS y AFP ofreció mayor liquidez a los peruanos lo que permitió en muchos casos que optaran por adquirir una vivienda propia, los créditos hipotecarios con tasas históricamente bajas, y los descuentos que están ofreciendo las entidades del sector inmobiliario forman parte importante de este repunte, además del impulso del Estado en políticas que faciliten la adquisición de viviendas, como el Bono Familiar Habitacional, Bono MiVivienda, MiVivienda Verde y Techo Propio. (blog UTP).

Según estudios realizados por la propia empresa, uno de los distritos con mayor oferta de departamentos es el distrito de Miraflores, dentro del sector Lima Top,

por ser el distrito más céntrico y turístico de Lima Metropolitana, con mayor seguridad ciudadana y da un mayor estatus económico.

La empresa realizó un estudio macroeconómico previo para identificar las variables a considerar para la estimación de las ventas, para ello, utilizó información propia y la data abierta de las series estadísticas mensuales del Banco de Reserva Central del Perú (BCRP) del período enero de 2017 hasta abril de 2022. Las variables consideradas fueron la siguientes: unidades en oferta, ventas de inmuebles, precios de oferta, precio de ventas, mes de feria inmobiliaria, tasa de interés de fondeo, tasa de interés activa, índice de precios inmueble, índice de precios por $m^2$ en distritos medios, índice de precios por m2 en distritos altos, índice de precios por $m^2$ en 12 distritos, índice de precios de consumidor Lima metropolitana alimentos & energía, producción de energía de Lima, importación de bienes de consumo duradero, índice de coyuntura de energía, índice de coyuntura de energía sin minas, índice de coyuntura de consumo de energía, variación porcentual demanda interna, variación porcentual del PBI, variación porcentual del PBI - sector construcción, índice de precios de inflación subyacente de Bienes, importaciones de materiales de construcción, índice de precios de importaciones, expectativa de PBI, tipo de cambio bancario mensual, indicador de variación mensual del tipo de cambio.

En el estudio de la estimación de las ventas de departamentos del distrito de Miraflores se consideraron las siguientes variables: índices de precios de importaciones. Índice de precios del consumidor en Lima (energía y alimentos), índice de precios de inmuebles, oferta total, tipo de cambio bancario mensual; además, el PBI se tomará en cuenta para verificar su influencia en las ventas.

Para estimar las futuras ventas de inmuebles con las 6 variables escogidas se usó el método de Ensamble por medias.

El término ensamblado estadístico (también denominado como fusión o combinación estadística) se refiere al conjunto de metodologías que permiten combinar varios modelos estadísticos para dar lugar a un meta-algoritmo que mejore los resultados de los modelos individuales que lo forman. Desde las

técnicas más sencillas, como el uso de promedios o medias, hasta las más sofisticadas, como bagging o boosting.
La metodología de Ensamble por medias usado para el proyecto, consiste de 4 pasos: Feature Engineering, Modelo ARIMA, Modelo de regresión lineal múltiple y por último la combinación de resultados de los modelos por la media simple.

El Feature Engineering es una técnica de aprendizaje automático que aprovecha los datos para crear nuevas variables que no están en el conjunto de entrenamiento, con el objetivo de simplificar y acelerar las transformaciones de datos al mismo tiempo, mejorar la precisión del modelo.

El Modelo ARIMA es una técnica de predicción cuantitativa estacionaria. Son modelos paramétricos que tratan de obtener la representación de la serie en términos de la interrelación temporal de sus elementos. El instrumento fundamental a la hora de analizar las propiedades de una serie temporal en términos de la interrelación temporal de sus observaciones es el denominado coeficiente de autocorrelación que mide la correlación, es decir, el grado de asociación lineal que existe entre observaciones separadas k periodos. Estos coeficientes de autocorrelación proporcionan mucha información sobre cómo están relacionadas entre sí las distintas observaciones de una serie temporal, lo que ayudará a construir el modelo apropiado para los datos. Por otro lado, proporcionan también información para predecir.

La Regresión Lineal Múltiple nos permite establecer la relación que se produce entre una variable dependiente (Y) y un conjunto de variables independientes (X1, X2, ... Xk). El análisis de regresión lineal múltiple, a diferencia del simple, se aproxima más a situaciones de análisis real puesto que los fenómenos, hechos y procesos sociales, por definición, son complejos y, en consecuencia, deben ser explicados en la medida de lo posible por la serie de variables que, directa e indirectamente, participan en su concreción.

Por último, el Ensamble por medias consistió en promediar los resultados obtenidos de la estimación de las ventas de inmuebles de los modelos

## 1.2. OBJETIVOS

### 1.2.1. Objetivo General

Determinar las ventas estimadas de departamentos del distrito de Miraflores de una empresa inmobiliaria de Lima utilizando Ensamble por Medias.

### 1.2.2. Objetivos Específicos

- Determinar las ventas estimadas de departamentos en las zonas 1, 2 y 3 del distrito de Miraflores.
- Determinar la robustez del modelo mediante el valor de $R^2$
- Determinar la periodicidad mensual o trimestral del modelo mediante la inclusión de la variable ratio promedio de 3 a 6 meses en la estimación de los modelos.

# II. REVISIÓN DE LA LITERATURA

## 2.1. Análisis de Series de Tiempo

### 2.1.1. Introducción

En muchas investigaciones estadísticas, el tipo de información disponible se concreta en observaciones recogidas en intervalos de tiempo (meses, trimestres, años, etc.) dicho conjunto de observaciones o datos a lo largo del tiempo forman lo que se denomina serie temporal.

### 2.1.2. Elementos de Ecuaciones en Diferencia

La ecuación en diferencia sirve para denominar un proceso similar o equivalente dentro de las ecuaciones diferenciales, dentro del cual se consideran a un conjunto de variables que están en función del tiempo. Así, si consideramos al tiempo como una variable continua, es decir, consideramos una variable Z(t), se puede expresar de manera siguiente la ecuación diferencial:

$$\frac{dZ(t)}{dt}; \frac{d^2Z(t)}{dt^2}; \ldots; \frac{d^kZ(t)}{dt^k} \qquad (3.1)$$

Por otro lado, suponiendo el caso del tiempo en forma discreta, con t = . . . , −2, −1, 0, 1, 2, . . ., entonces el comportamiento de la serie de variables dadas por Zt , la cual se puede expresar como:

$$\Delta Z_t; \Delta^2 Z_t; \ldots; \Delta^k Z_t \qquad (3.2)$$

Observemos que una forma técnicamente más correcta es escribir las expresiones anteriores como:

$$\frac{\Delta Z_t}{\Delta t}; \frac{\Delta^2 Z_t}{\Delta t^2}; \ldots; \frac{\Delta^k Z_t}{\Delta t^k} \qquad (3.3)$$

No obstante, no pasa desapercibido que Δt = 1, por lo que resultan equivalentes ambos conjuntos de expresiones (3.2) y (3.3).

### 2.1.3. Clasificación

Para la metodología de estudio, se presentará el modelo de series de tiempo de Box y Jenkis (1970), siendo de importancia en esta metodología los Procesos Autorregresivos (AR) y de Medias móviles (MA), además de la combinación de ambas.

**Procesos Autorregresivos**

Los procesos autorregresivos tienen su origen en el trabajo de Cochrane y Orcutt de 1949, mediante el cual analizaron los residuales de una regresión clásica como un proceso autorregresivo. Su proceso se representa como una suma ponderada de observaciones pasadas de la variable. El número de rezagos (p) determina el orden del modelo autorregresivo.
El modelo en general se describe de la siguiente manera:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + a_t$$

Dónde: t=1,2,3, …

**Proceso de Medias Móviles**

El modelo de medias móviles de orden finito q, MA(q), es una aproximación natural al modelo lineal general. Se obtiene un modelo finito por el simple procedimiento de truncar el modelo de medias móviles de orden infinito.

$$Y_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \qquad a_t \sim RB(0, \sigma^2)$$

El número de rezagos del error considerados (q) determina el orden del modelo de media móvil.

**Procesos Arma y Arima**

**Modelo Arma:**

Los modelos presentados (AR, MA y ARMA) tienen una característica común: son usados para representar procesos estocásticos estacionarios, es decir, que las variables están ordenadas cronológicamente y que la distribución de probabilidad de cada observación permanece constante en el tiempo.
Ejemplo modelo ARMA (1,1)

$$y_t = a + by_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$$

**Modelo Arima:**

Muchas series de tiempo no son estacionarias, por ejemplo, el Producto Nacional Bruto o la Producción Industrial. Un tipo especial de series no estacionarias, son las no estacionarias homogéneas que se caracterizan porque, al ser diferenciadas una o más veces, se vuelven estacionarias.

- $\Delta Y_t = Y_t - Y_{t-1}$
- $\Delta^{n+1} Y_t = \Delta^n Y_t - \Delta^n Y_{t-1}$

Si después de haber diferenciado la serie Yt se consigue una serie estacionaria Wt, y dicha serie obedece a un proceso ARMA (p, q), se dice que Yt responde a un proceso ARIMA (p, d, q).

Para la correcta identificación del modelo ARIMA es necesario:

- Determinar el grado de homogeneidad u orden de integración de la serie
- Determinar el orden de las partes de promedio móvil y autorregresivas
- Evaluar los distintos modelos construidos

**Función de Autocorrelación Parcial**

La autocorrelación parcial identifica la relación entre los valores actuales y los valores anteriores de la serie cronológica original, después de quitar los efectos de las autocorrelaciones de orden inferior. El correlograma PACF de autocorrelaciones parciales se utiliza debido a la relación entre las observaciones.

**Criterios:**

- Si ninguna de las autocorrelaciones es significativamente diferente de cero, la serie es esencialmente ruido blanco.
- Si las autocorrelaciones decrecen linealmente, pasando por el cero, o muestra un patrón cíclico, pasando por cero varias veces, la serie no es estacionaria.
- Si las autocorrelaciones muestran estacionalidad, o se tiene un alza cada periodo la serie no es estacionaria y hay que diferenciarla con un salto igual al periodo.
- Si las autocorrelaciones decrecen exponencialmente hacia cero y las autocorrelaciones parciales son significativamente no nulas sobre un pequeño número de rezagos, se puede usar un modelo autorregresivo
- Si las autocorrelaciones parciales decrecen exponencialmente hacia cero y las autocorrelaciones son significativamente no nulas sobre un pequeño número de rezagos, se puede usar un modelo de medias móviles
- Si las autocorrelaciones simples y parciales decrecen lentamente hacia cero, pero sin alcanzar el cero, se puede usar un modelo autorregresivo combinado con medias móviles

**Pronósticos**

Todo pronóstico tiene asociado un alcance, pudiendo ser éste de corto, mediano o largo plazo. Los horizontes de tiempo correspondientes a dichos alcances dependerán de la industria bajo estudio. En cuanto al atractivo de uno u otro pronóstico, Éste estará sujeto al tipo de decisión que se desea tomar o de acción en desarrollo. A modo de ejemplo, en la industria del cobre, alcances convencionales y decisiones comunes en el mercado y la industria son:

**Tabla 1.**

*Tipos de pronóstico*

| Tipo de pronóstico | Alcance | Decisiones |
|---|---|---|
| Cortísimo Plazo | Minutos, horas | Operaciones especulativas |
| Corto Plazo | Días, semanas, meses, un año | Operaciones especulativas, de cobertura y de gestión comercial |
| Mediano Plazo | Uno a seis años | Evaluación y control de los resultados de la gestión y de los negocios de una empresa |
| Largo Plazo | 6 a 50 años | Planificación de la producción y evaluación de proyectos |

***Nota:*** Elaborado en base al libro series de tiempo, universidad de Chile (2008).

A mayor alcance del pronóstico, mayor será el nivel de incertidumbre que se alcanzará.

## 2.2. Análisis de Regresión Lineal Múltiple

### 2.2.1. Definición

El Análisis de Regresión Lineal Múltiple nos permite establecer la relación que se produce entre una variable dependiente Y y un conjunto de variables independientes (X1, X2, ... XK). El análisis de regresión lineal múltiple, a diferencia del simple, se aproxima más a situaciones de análisis real puesto que los fenómenos, hechos y procesos sociales, por definición, son complejos y, en consecuencia, deben ser explicados en la medida de lo posible por la serie de variables que, directa e indirectamente, participan en su concreción.

### 2.2.2. Supuestos

Entre los supuestos para la regresión lineal múltiple tenemos:

- La distribución de probabilidad de error es normal
- La varianza de la distribución del error es constante para todos los X,
- La media de la distribución de probabilidad del error es 0.
- Los valores de error son independientes entre sí. Esta suposición indica que se ha elegido una muestra aleatoria de objetos a partir de la población para medirlos.

### 2.2.3. Ecuación de Regresión

En el modelo de regresión lineal múltiple esperamos que los sucesos tengan una forma funcional como:

$$y_j = b_o + b_1 x_{1j} + b_2 x_{2j} + \cdots + b_k x_{kj} + u_j$$

donde y es la variable endógena, x las variables exógenas, u los residuos y b los coeficientes estimados del efecto marginal entre cada x e y.

### 2.2.4. Prueba de Hipótesis

Las pruebas de hipótesis incluyen el uso de evidencia muestral para evaluar la probabilidad de que una suposición sobre alguna característica de una población sea cierta. El objetivo es encontrar la mejor ecuación para predecir Y y después decidir si esta ecuación satisface lo que se desea probar.

**Coeficiente de Determinación Múltiple:**

Mide la proporción de la variabilidad de la variable dependiente explicada por las variables independiente que en ese momento han sido admitidas en el modelo. A partir del resumen de los modelos generados paso a paso podemos calcular el incremento de R2, siendo éste una estimación de la importancia relativa que tiene

la variable que acabamos de introducir en el paso correspondiente para predecir la variable dependiente.

Un valor de R2 cercano a 1, significa que la ecuación de regresión es muy exacta porque explica una gran proporción de la variabilidad de Y.

La ecuación de cálculo de R2 para la regresión múltiple se determina de la siguiente manera:

$$R^2 = 1 - \frac{\sum\left(y - \hat{y}\right)^2}{\sum\left(y - \bar{y}\right)^2}$$

### 2.2.5. Selección de Variables

En el análisis de regresión lineal múltiple la construcción de su correspondiente ecuación se realiza seleccionando las variables una a una, "paso a paso". La finalidad perseguida es buscar de entre todas las posibles variables explicativas aquellas que más y mejor expliquen a la variable dependiente sin que ninguna de ellas sea combinación lineal de las restantes. Este procedimiento implica que: (1) en cada paso solo se introduce aquella variable que cumple unos criterios de entrada; (2) una vez introducida, en cada paso se valora si alguna de las variables cumple criterios de salida; y (3), en cada paso se valora la bondad de ajuste de los datos al modelo de regresión lineal y se calculan los parámetros del modelo verificado en dicho paso. El proceso se inicia sin ninguna variable independiente en la ecuación de regresión y el proceso concluye cuando no queda ninguna variable fuera de la ecuación que satisfaga el criterio de selección (garantiza que las variables seleccionadas son significativas) y/o el criterio de eliminación (garantizar que una variable seleccionada no es redundante).

Verificación de los criterios de probabilidad de entrada. El p-valor asociado al estadístico T, o probabilidad de entrada, nos indica si la información proporcionada por cada una de las variables es redundante. Si éste es menor que

un determinado valor crítico, la variable será seleccionada. El SPSS por defecto establece en 0.05 el valor crítico de la probabilidad de entrada. El criterio de tolerancia puede ser aplicado como un criterio adicional a la probabilidad de entrada. Éste nos ayuda a identificar si alguna de las variables del modelo es una combinación lineal de las restantes. Si dicho valor es próximo a 0, la variable analizada será una combinación lineal de las restantes variables independientes introducidas. Si el valor de la tolerancia se aproxima a 1 puede reducir la parte de la variabilidad de Y no explicada por las restantes. En síntesis, si la tolerancia para una variable es muy pequeña se excluirá del modelo.

Verificación del criterio de probabilidad de salida. En este caso, si el p-valor asociado al estadístico T, o probabilidad de salida, es mayor que un determinado valor crítico, la variable será eliminada. En el caso práctico que recogemos en los resultados puede apreciarse que las dos variables independientes han superado los criterios de entrada y de salida.

Límite al número de pasos. Para evitar que el proceso de selección se convierta en un proceso cíclico se debe establecer un número límite de pasos. Normalmente este límite es el que equivale al doble del número de variables independientes.

## 2.3. Modelo de Ensamble

### 2.3.1. Definición

El término ensamblado estadístico se refiere al conjunto de metodologías que permite combinar varios modelos estadísticos para dar lugar a un meta-algoritmo que mejore los resultados de los modelos individuales que lo forman.

### 2.3.2. Motivación

**Razones Estadísticas:**

Cualquier algoritmo de aprendizaje estadístico se puede ver como la búsqueda del mejor modelo para un conjunto de datos. Sea cual sea el procedimiento para escoger un modelo y sus parámetros, siempre existirá una incertidumbre asociada al proceso debido a que se dispone sólo de un conjunto de entrenamiento finito.

**Razones Computacionales:**

Algunos algoritmos, como las redes neuronales o los árboles de decisión, pueden dar lugar a minimizar funciones de coste altamente no convexas, pudiendo quedar los métodos utilizados para resolverlas en un óptimo local. Ensamblar distintas hipótesis, utilizando en cada una de ellas puntos de partida distintos para esa búsqueda local, puede aumentar la probabilidad de aproximar mejor la hipótesis verdadera.

**Razones Representacionales:**

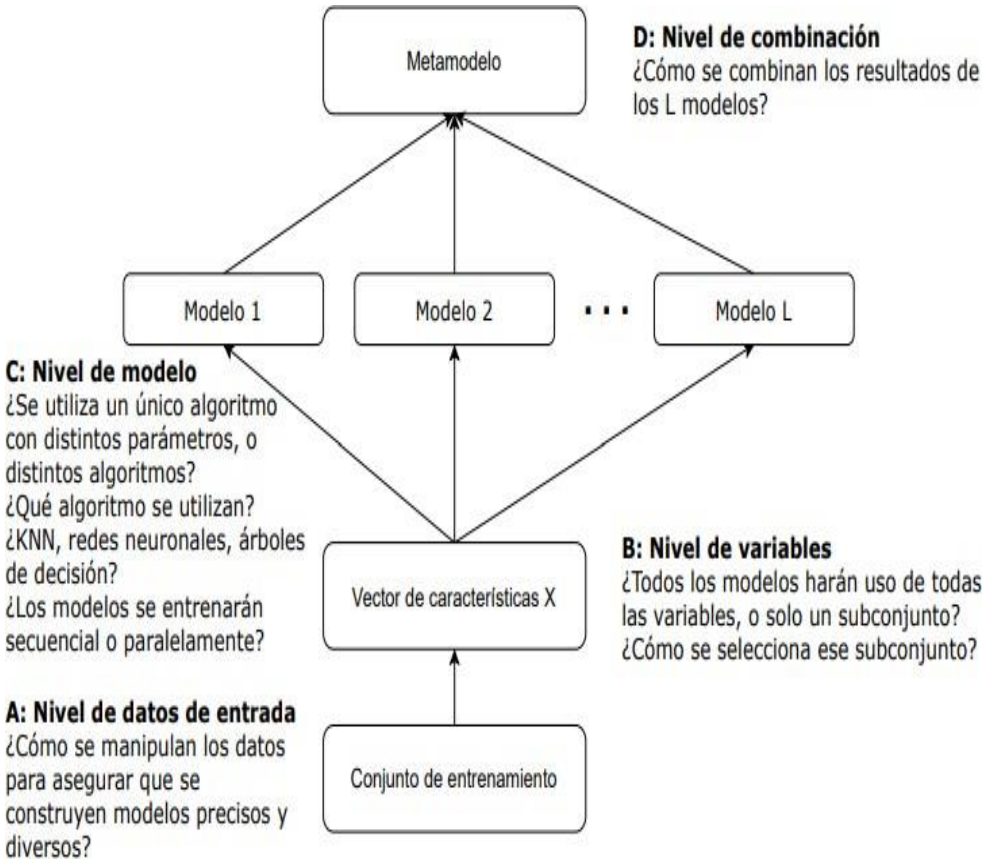
En algunos casos, la función verdadera que se quiere aproximar no puede ser representada por ninguno de los modelos del espacio de hipótesis. Por ejemplo, un ensamblado de clasificadores lineales puede dar lugar a fronteras de decisión tan complejas como se quiera, dando así la posibilidad de resolver un problema no lineal no factible en caso de usar un único clasificador lineal.

### 2.3.3. Técnicas de Ensamble

Kuncheva (2014) propone la siguiente metodología de ensamble:

**Figura 1**

*Metodología de ensamble propuesto por Kuncheva (2014).*




*Nota.* Tomado del trabajo en Métodos de ensamblado en Machine Learning (p. 23), por Ricarey Fernández, Ricardo, 2021.

Entre los métodos de ensamble conocidas tenemos: promedios, promedio bayesiano, nieve bayes ensemble, bagging, random forest, boosting, stacking y más.

# III. DESARROLLO DEL TRABAJO

## 3.1. Delimitación Temporal y Geográfica

El estudio se realizó en el 2022, con información de datos abiertos del BCRP y base de datos de una empresa inmobiliaria de Lima.

## 3.2. Naturaleza del Trabajo

El presente proyecto fue realizado para conocer el futuro de las ventas del mercado inmobiliario a través del conocimiento estadístico aplicado, los cuales ayudó a comprender lo cambiante que son los distintos factores influyentes en el Sector. Aplicar el conocimiento adecuado de los procedimientos de las técnicas usadas, hubiera correspondido a dar continuidad al estudio y enriquecer el conocimiento futuro del sector a los distintos agentes participantes del rubro.

### 3.2.1. Enfoque, Tipo y Diseño de la Investigación

El enfoque de la investigación es aplicado, porque su principal motivación es el deseo de descubrir nuevos conocimientos, es básica porque sirve de cimiento para realizar la investigación aplicada, y es fundamental porque es esencial para el desarrollo de la ciencia.

El tipo de la investigación es explicativo, siendo el propósito estimar las ventas de departamentos en el distrito de Miraflores (variable dependiente) a través de los índices de precios de importaciones, el índice de precios del consumidor en Lima (energía y alimentos), el índice de precios de inmuebles, la oferta total, el tipo de cambio bancario mensual y el PBI (variables independientes).

El diseño de la investigación es No Experimental, no existe manipulación en las variables independientes y no hay control total de las fuentes de validez interna.

Es de carácter transversal porque la recolección de la información se realiza en un momento determinado del tiempo.

### 3.2.2. Variables de la Investigación

**Variable dependiente:**

Y: Ventas de inmuebles del distrito de Miraflores

**Variables independientes:**

X1: Índices de precios de importaciones
X2: Índice de precios del consumidor en Lima (energía y alimentos),
X3: Índice de precios de inmuebles,
X4: Oferta total
X5: Tipo de cambio bancario mensual
X6: Expectativas del PBI

### 3.2.3. Población y Muestra

La población para construir el modelo fueron los 63 registros de ventas de departamentos del distrito de Miraflores correspondiente al período de enero del 2017 hasta abril del 2022. No se consideró una muestra representativa porque el estudio se realizó a nivel poblacional.

### 3.2.4. Técnicas e Instrumentos de Recolección de Datos

La técnica de recolección de datos fue la observación de la base de datos para identificar los registros de ventas mensuales de departamentos en el distrito de Miraflores correspondiente al período de enero de 2017 hasta abril de 2022. El instrumento de recolección de datos fue la ficha de registros.

### 3.2.5. Técnicas de Procesamiento y Análisis de Datos

La técnica de procesamiento de datos se realizó a través del software Python y la técnica de análisis de datos el Modelo de Ensamble por la Media Simple.

## 3.3. Contribución en la Solución Problemática

Con el trabajo de suficiencia profesional se logró resolver la problemática laboral de la empresa inmobiliaria de lograr estimar las ventas de departamentos del distrito de Miraflores y de las 3 zonas de análisis de estudio. Para ello, se utilizó en la primera etapa las técnicas estadísticas de Series de Tiempo y Análisis de Regresión Lineal Múltiple y en la segunda etapa el Método de Ensamble por Medias. Previamente, se identificaron las variables macroeconómicas intervinientes que influyen en las ventas de inmuebles, descartando aquellas consideradas como no influyentes.

## 3.4. Contribución de acuerdo a competencias y habilidades adquiridas

En la formación recibida en la carrera profesional de Estadística Informática de la Universidad Nacional Agraria La Molina se consideraron un conjunto de cursos del área de las técnicas multivariadas. El aprendizaje de estas materias permitió identificar las técnicas estadísticas de Series de Tiempo y Análisis de Regresión Lineal Múltiple para el análisis de los datos en la fase preliminar y el Modelo de Ensamble por Medias como el más adecuado para resolver la problemática laboral. Para el desarrollo de este estudio se realizó una búsqueda bibliográfica para revisar diferentes modelos estadísticos que fueran pertinentes para obtener resultados más robustos y para estimar las variables a predecir, tal como fue en la estimación de las futuras ventas de inmuebles.

## 3.5. Beneficio del Centro laboral por Contribución en la Solución

Si bien la continuidad del estudio no fue lo esperado, esto implicaba un mayor coste de investigación que no compensaba con la distribución del resultado a los

clientes potenciales interesados en el estudio. Al no conocer la totalidad del proceso de modelización, se decidió por dejar en stand by el estudio y continuar con otros productos de data para las ventas de inmuebles, la cual apertura la posibilidad de seguir buscando otras metodologías que impliquen un menor costo a partir de la experiencia obtenida en el estudio.

# IV. RESULTADOS Y DISCUSIÓN

El método de Ensamble por Medias utilizado en el estudio toma en consideración tres etapas. En la Etapa 1, se aplica el modelo ARIMA basado en el aprendizaje de la misma serie temporal tomando como input la misma serie y su comportamiento histórico. En la Etapa 2, se utiliza el modelo de regresión lineal para que el sistema aprenda de las interacciones y elasticidades que tienen los componentes macroeconómicos sobre la venta de inmuebles. Finalmente, en la Etapa 3, se ensamblan los dos modelos desarrollados en las etapas 1 y 2. Para fines del estudio, se dividió el total de ventas de inmuebles en Miraflores en 3 zonas.

## Miraflores Zona 1

### Etapa 1

Se utiliza la serie de tiempo ARIMA y se determinan sus componentes p, d y q mediante la prueba de Dicky-Fuller de autoregresión y estacionariedad.

.

**Figura 2.**

*Prueba de Dicky Fuller para estacionariedad*

```
result = adfuller(X_train[target].dropna())
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
```

```
ADF Statistic: -3.962214
p-value: 0.001622
```

*Nota*: Resultado de script para evaluar la estacionariedad

El valor de $p < 0.05$ indica que con 95% de confianza se rechaza el Ho, es decir la serie temporal es no estacionaria lo que significa que ser necesario modelar la serie con un enfoque ARIMA.

**Figura 3**

*Resultado de modelo ARIMA con valor p,d,q (1,2,1)*

ARIMA Model Results

| | | | |
|---|---|---|---|
| Dep. Variable: | D2.Venta_zona_1 | No. Observations: | 43 |
| Model: | ARIMA(1, 2, 1) | Log Likelihood | -143.792 |
| Method: | css-mle | S.D. of innovations | 6.517 |
| Date: | Mon, 03 Oct 2022 | AIC | 297.585 |
| Time: | 01:25:48 | BIC | 306.391 |
| Sample: | 2 | HQIC | 300.832 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 1.0562 | 1.257 | 0.840 | 0.401 | -1.408 | 3.520 |
| Venta_zona_1_lag2_avg3 | -0.0840 | 0.102 | -0.824 | 0.410 | -0.284 | 0.116 |
| ar.L1.D2.Venta_zona_1 | -0.2812 | 0.148 | -1.894 | 0.058 | -0.572 | 0.010 |
| ma.L1.D2.Venta_zona_1 | -1.0000 | 0.064 | -15.505 | 0.000 | -1.126 | -0.874 |

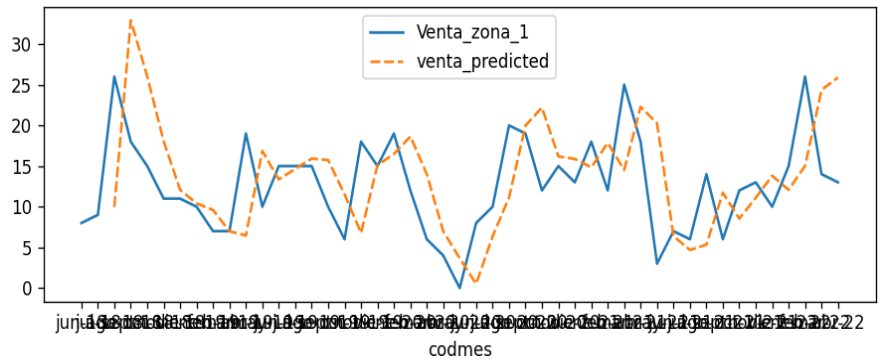
Roots

| | Real | Imaginary | Modulus | Frequency |
|---|---|---|---|---|
| AR.1 | -3.5561 | +0.0000j | 3.5561 | 0.5000 |
| MA.1 | 1.0000 | +0.0000j | 1.0000 | 0.0000 |

*Nota.* Resultado de script para estimar modelo ARIMA

**Figura 4**

*Gráfico de data original versus datos predichos por el Modelo ARIMA*




**Etapa 2**

Para el modelo de regresión, se eligió por experticia las variables significativas

**Figura 5**

*Correlación entre variables dependientes versus ventas*

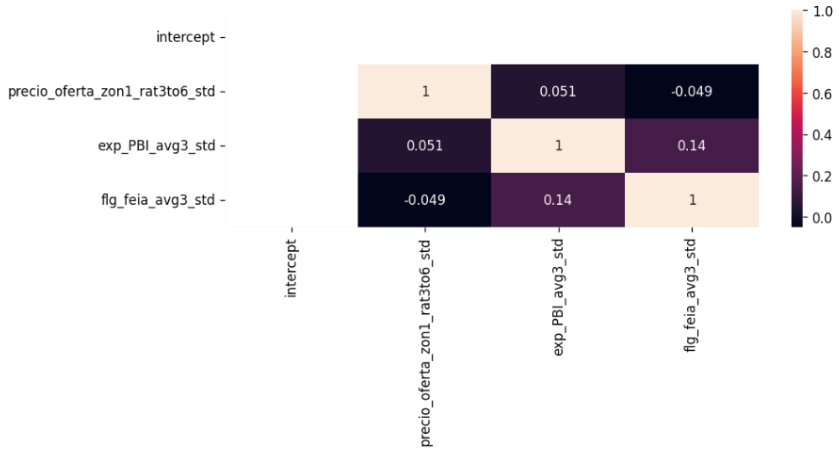



**Figura 6**

*Resultados de modelo de regresión*

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Venta_zona_1 | R-squared: | 0.307 |
| Model: | OLS | Adj. R-squared: | 0.253 |
| Method: | Least Squares | F-statistic: | 5.751 |
| Date: | Mon, 03 Oct 2022 | Prob (F-statistic): | 0.00234 |
| Time: | 01:34:06 | Log-Likelihood: | -129.54 |
| No. Observations: | 43 | AIC: | 267.1 |
| Df Residuals: | 39 | BIC: | 274.1 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| intercept | 12.7242 | 0.789 | 16.135 | 0.000 | 11.129 | 14.319 |
| precio_oferta_zon1_rat3to6_std | -2.2126 | 0.788 | -2.809 | 0.008 | -3.806 | -0.619 |
| exp_PBI_avg3_std | 1.2573 | 0.793 | 1.586 | 0.121 | -0.346 | 2.861 |
| flg_feia_avg3_std | 1.8231 | 0.789 | 2.312 | 0.026 | 0.228 | 3.418 |

| | | | |
|---|---|---|---|
| Omnibus: | 3.670 | Durbin-Watson: | 1.614 |
| Prob(Omnibus): | 0.160 | Jarque-Bera (JB): | 3.280 |
| Skew: | 0.671 | Prob(JB): | 0.194 |
| Kurtosis: | 2.834 | Cond. No. | 1.18 |

*Nota.* Resultado de script para estimar modelo de Regresión Múltiple

Debido a que algunos p-values superan los valores de 0.05 (confianza de 95%) se consideró adecuado considerar algunos componentes para mejorar la robustez que puede presentar el modelo frente a cambios sistémicos del país.

## Etapa 3

La combinación de modelos en esta zona se realizó bajo la metodología de ensamble por medias.

**Tabla 2**

*Resultados de ensamble de zona 1 de Miraflores*

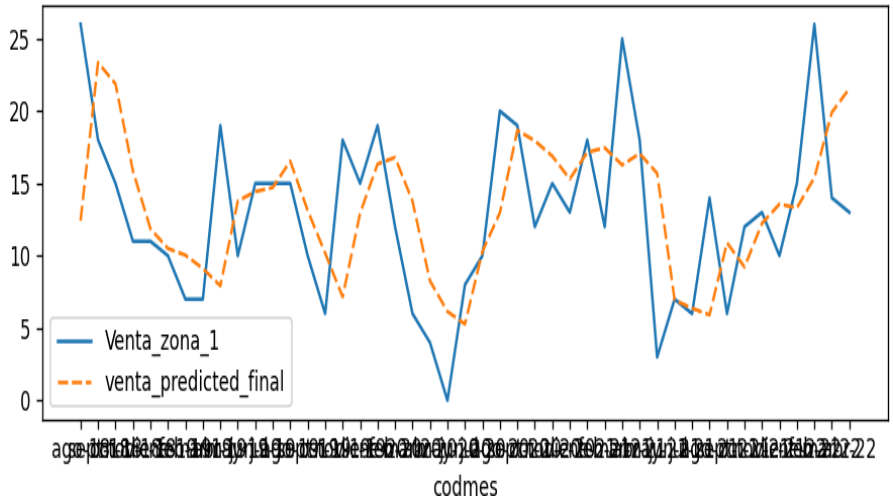
| codmes | Venta_zona_1 | venta_predicted | macro_pred | venta_predicted_final |
|---|---|---|---|---|
| **ago-18** | 26 | 9.96426 | 14.881176 | 12.422718 |
| **sep-18** | 18 | 32.915671 | 13.779123 | 23.347397 |
| **oct-18** | 15 | 26.02703 | 17.707776 | 21.867403 |
| **nov-18** | 11 | 18.055119 | 13.64571 | 15.850414 |
| **dic-18** | 11 | 12.073504 | 11.658129 | 11.865816 |
| **ene-19** | 10 | 10.412365 | 10.594728 | 10.503546 |
| **feb-19** | 7 | 9.601564 | 10.478188 | 10.039876 |
| **mar-19** | 7 | 6.985857 | 11.259153 | 9.122505 |
| **abr-19** | 19 | 6.461188 | 9.381773 | 7.921481 |
| **may-19** | 10 | 16.858914 | 10.702949 | 13.780931 |
| **jun-19** | 15 | 13.355183 | 15.510753 | 14.432968 |
| **jul-19** | 15 | 14.645956 | 14.781013 | 14.713485 |
| **ago-19** | 15 | 15.91803 | 17.148844 | 16.533437 |
| **sep-19** | 10 | 15.737176 | 10.50691 | 13.122043 |
| **oct-19** | 6 | 11.527562 | 8.840388 | 10.183975 |
| **nov-19** | 18 | 6.770102 | 7.553218 | 7.16166 |
| **dic-19** | 15 | 15.119152 | 10.762849 | 12.941001 |
| **ene-20** | 19 | 16.487715 | 16.184714 | 16.336215 |
| **feb-20** | 12 | 18.648171 | 14.933299 | 16.790735 |
| **mar-20** | 6 | 13.991843 | 13.543828 | 13.767835 |
| **abr-20** | 4 | 6.977804 | 9.552613 | 8.265209 |
| **may-20** | 0 | 3.669943 | 8.629879 | 6.149911 |
| **jun-20** | 8 | 0.515633 | 10.036335 | 5.275984 |
| **jul-20** | 10 | 6.358791 | 14.238453 | 10.298622 |
| **ago-20** | 20 | 11.132815 | 14.818031 | 12.975423 |
| **sep-20** | 19 | 19.984804 | 17.341451 | 18.663128 |
| **oct-20** | 12 | 22.188595 | 13.692533 | 17.940564 |
| **nov-20** | 15 | 16.188278 | 17.579429 | 16.883853 |
| **dic-20** | 13 | 15.874056 | 14.725101 | 15.299579 |
| **ene-21** | 18 | 14.846618 | 19.440002 | 17.14331 |
| **feb-21** | 12 | 17.852423 | 17.060817 | 17.45662 |
| **mar-21** | 25 | 14.51065 | 18.028458 | 16.269554 |
| **abr-21** | 18 | 22.276894 | 11.841405 | 17.05915 |
| **may-21** | 3 | 20.248401 | 11.138525 | 15.693463 |
| **jun-21** | 7 | 6.379265 | 7.500588 | 6.939927 |

| | | | | |
|---|---|---|---|---|
| **jul-21** | 6 | 4.685962 | 8.05645 | 6.371206 |
| **ago-21** | 14 | 5.335036 | 6.501954 | 5.918495 |
| **sep-21** | 6 | 11.719623 | 10.087367 | 10.903495 |
| **oct-21** | 12 | 8.54195 | 9.895144 | 9.218547 |
| **nov-21** | 13 | 11.10509 | 13.363458 | 12.234274 |
| **dic-21** | 10 | 13.810417 | 13.322359 | 13.566388 |
| **ene-22** | 15 | 12.076891 | 14.537586 | 13.307238 |
| **feb-22** | 26 | 15.02508 | 15.756459 | 15.390769 |
| **mar-22** | 14 | 24.351519 | 15.46013 | 19.905825 |
| **abr-22** | 13 | 25.886021 | 17.211175 | 21.548598 |

***Nota:*** Contenido de resultado de script para elaboración de estudio.

**Figura 7**

*Gráfica de modelo de ensamble*




***Miraflores Zona 2***

**Etapa 1**

Se aplica la serie de tiempo ARIMA y para determinar sus componentes p, d y q se utiliza la prueba de Dicky Fuller de autoregresión y estacionariedad.

**Figura 8**

*Prueba de Dicky Fuller para estacionariedad*

```
result = adfuller(X_train[target].dropna())
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
```

```
ADF Statistic: -1.770886
p-value: 0.395009
```

*Nota*: Resultado de script para evaluar la estacionariedad

El valor de p > 0.05 indica que con 95% de confianza no se rechaza el Ho, es decir la serie temporal es estacionaria no es necesario modelar la serie con un enfoque ARIMA.

**Figura 9**

*Resultado de modelo ARIMA con valor p,d,q (1,2,1)*

ARIMA Model Results

| | | | |
|---|---|---|---|
| Dep. Variable: | D2.Venta_zona_2 | No. Observations: | 43 |
| Model: | ARIMA(1, 2, 1) | Log Likelihood | -172.267 |
| Method: | css-mle | S.D. of innovations | 12.663 |
| Date: | Mon, 03 Oct 2022 | AIC | 354.534 |
| Time: | 01:59:50 | BIC | 363.340 |
| Sample: | 2 | HQIC | 357.781 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.0173 | 0.839 | -0.021 | 0.984 | -1.662 | 1.627 |
| Venta_zona_2_lag2_avg3 | 0.0021 | 0.022 | 0.096 | 0.923 | -0.040 | 0.044 |
| ar.L1.D2.Venta_zona_2 | -0.1969 | 0.151 | -1.302 | 0.193 | -0.493 | 0.099 |
| ma.L1.D2.Venta_zona_2 | -1.0000 | 0.062 | -16.210 | 0.000 | -1.121 | -0.879 |

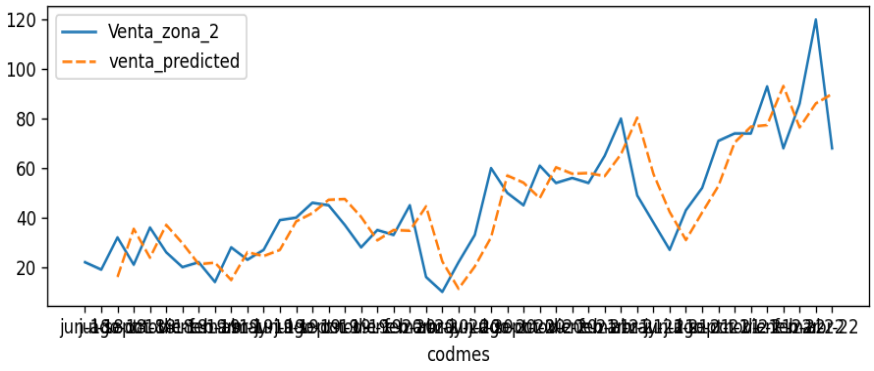
Roots

| | Real | Imaginary | Modulus | Frequency |
|---|---|---|---|---|
| AR.1 | -5.0797 | +0.0000j | 5.0797 | 0.5000 |
| MA.1 | 1.0000 | +0.0000j | 1.0000 | 0.0000 |

*Nota.* Resultado de script para estimar modelo ARIMA

**Figura 10**

*Gráfico de data inicial versus datos predecidos por el Modelo ARIMA*




**Etapa 2**

Para el modelo de regresión, se eligió por expertis las variables significativas

**Figura 11**

*Correlación entre variables dependientes versus ventas*

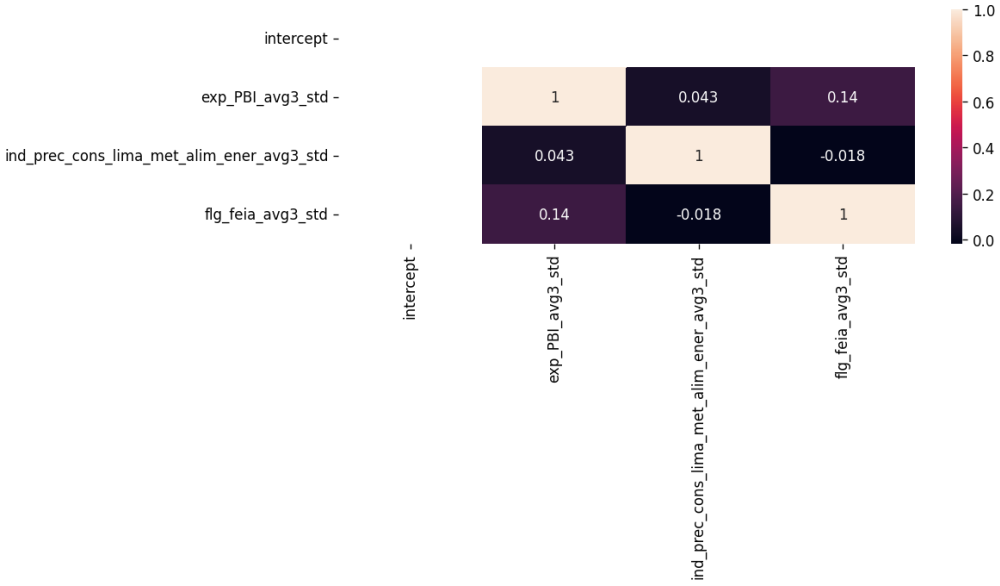
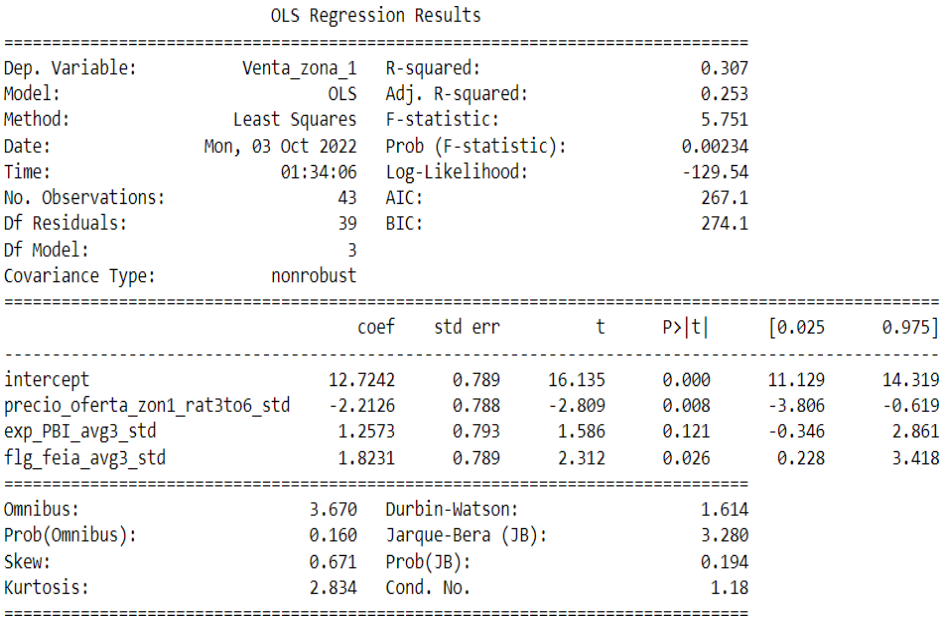


**Figura 12**

*Resultados de modelo de regresión*



OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Venta_zona_1 | R-squared: | 0.307 |
| Model: | OLS | Adj. R-squared: | 0.253 |
| Method: | Least Squares | F-statistic: | 5.751 |
| Date: | Mon, 03 Oct 2022 | Prob (F-statistic): | 0.00234 |
| Time: | 01:34:06 | Log-Likelihood: | -129.54 |
| No. Observations: | 43 | AIC: | 267.1 |
| Df Residuals: | 39 | BIC: | 274.1 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| intercept | 12.7242 | 0.789 | 16.135 | 0.000 | 11.129 | 14.319 |
| precio_oferta_zon1_rat3to6_std | -2.2126 | 0.788 | -2.809 | 0.008 | -3.806 | -0.619 |
| exp_PBI_avg3_std | 1.2573 | 0.793 | 1.586 | 0.121 | -0.346 | 2.861 |
| flg_feia_avg3_std | 1.8231 | 0.789 | 2.312 | 0.026 | 0.228 | 3.418 |

| | | | |
|---|---|---|---|
| Omnibus: | 3.670 | Durbin-Watson: | 1.614 |
| Prob(Omnibus): | 0.160 | Jarque-Bera (JB): | 3.280 |
| Skew: | 0.671 | Prob(JB): | 0.194 |
| Kurtosis: | 2.834 | Cond. No. | 1.18 |

*Nota.* Resultado de script para estimar modelo de Regresión Múltiple

Debido a que algunos p-values superan los valores de 0.05 (confianza de 95%) se consideró conveniente tomar en cuenta algunos componentes para mejorar la robustez del modelo frente a cambios sistémicos del país.

## Etapa 3

Se utiliza el método de ensamblé de medias debido a que los vectores autorregresivos no estuvieron dentro de lo esperado a nivel de predicción.

**Tabla 3**

*Resultados de ensamble de zona 2 de Miraflores*

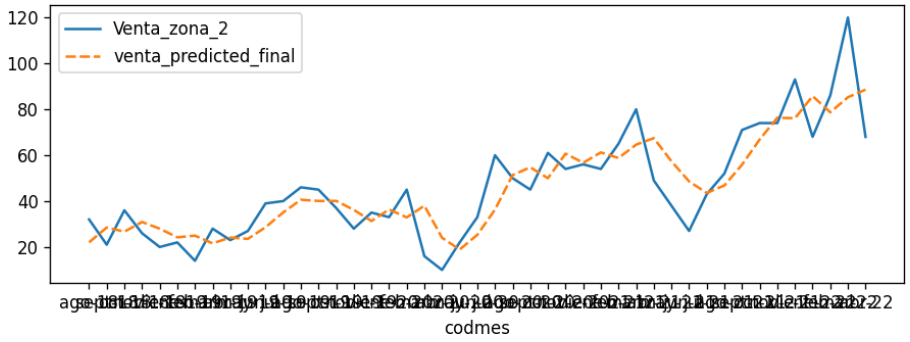
| codmes | Venta_zona_2 | venta_predicted | macro_pred | venta_predicted_final |
|---|---|---|---|---|
| **ago-18** | 32 | 16.031113 | 28.040954 | 22.036034 |
| **sep-18** | 21 | 35.474155 | 21.531488 | 28.502822 |
| **oct-18** | 36 | 23.819064 | 29.533635 | 26.676349 |
| **nov-18** | 26 | 37.057837 | 24.741473 | 30.899655 |
| **dic-18** | 20 | 29.662696 | 26.308425 | 27.985561 |
| **ene-19** | 22 | 21.220213 | 27.223059 | 24.221636 |

| | | | | |
|---|---|---|---|---|
| **feb-19** | 14 | 21.809536 | 28.09273 | 24.951133 |
| **mar-19** | 28 | 14.797577 | 28.406908 | 21.602242 |
| **abr-19** | 23 | 26.016506 | 22.207506 | 24.112006 |
| **may-19** | 27 | 24.476123 | 22.626769 | 23.551446 |
| **jun-19** | 39 | 26.973853 | 30.28074 | 28.627297 |
| **jul-19** | 40 | 38.47149 | 31.63151 | 35.0515 |
| **ago-19** | 46 | 41.809067 | 39.427943 | 40.618505 |
| **sep-19** | 45 | 47.196068 | 32.909026 | 40.052547 |
| **oct-19** | 37 | 47.504623 | 32.665491 | 40.085057 |
| **nov-19** | 28 | 40.299286 | 32.079743 | 36.189515 |
| **dic-19** | 35 | 30.844084 | 31.768521 | 31.306303 |
| **ene-20** | 33 | 35.002329 | 37.731106 | 36.366718 |
| **feb-20** | 45 | 34.730288 | 31.041605 | 32.885947 |
| **mar-20** | 16 | 44.555637 | 31.45815 | 38.006894 |
| **abr-20** | 10 | 22.316038 | 25.748668 | 24.032353 |
| **may-20** | 22 | 11.279638 | 26.477744 | 18.878691 |
| **jun-20** | 33 | 20.250214 | 30.354436 | 25.302325 |
| **jul-20** | 60 | 32.007843 | 40.539005 | 36.273424 |
| **ago-20** | 50 | 57.023469 | 45.653572 | 51.338521 |
| **sep-20** | 45 | 54.101424 | 55.370855 | 54.736139 |
| **oct-20** | 61 | 47.869778 | 52.06559 | 49.967684 |
| **nov-20** | 54 | 60.316244 | 61.022631 | 60.669437 |
| **dic-20** | 56 | 57.732091 | 55.664513 | 56.698302 |
| **ene-21** | 54 | 58.013229 | 64.422022 | 61.217626 |
| **feb-21** | 65 | 56.789244 | 60.800866 | 58.795055 |
| **mar-21** | 80 | 65.605381 | 63.595014 | 64.600197 |
| **abr-21** | 49 | 80.381265 | 54.520482 | 67.450873 |
| **may-21** | 38 | 57.645703 | 57.071939 | 57.358821 |
| **jun-21** | 27 | 42.282516 | 54.883665 | 48.583091 |
| **jul-21** | 43 | 30.975371 | 56.124311 | 43.549841 |
| **ago-21** | 52 | 42.068956 | 51.566211 | 46.817583 |
| **sep-21** | 71 | 52.780145 | 58.956303 | 55.868224 |
| **oct-21** | 74 | 70.361102 | 63.343857 | 66.85248 |
| **nov-21** | 74 | 76.713378 | 75.95168 | 76.332529 |
| **dic-21** | 93 | 77.375481 | 74.828504 | 76.101993 |
| **ene-22** | 68 | 93.167768 | 78.359143 | 85.763455 |
| **feb-22** | 86 | 76.415743 | 80.971964 | 78.693853 |
| **mar-22** | 120 | 86.125501 | 84.393558 | 85.25953 |
| **abr-22** | 68 | 89.952038 | 87.196582 | 88.57431 |

***Nota:*** Contenido de resultado de script para elaboración de estudio.

**Figura 13**

*Gráfica de modelo de ensamble*




## Miraflores Zona 3

### Etapa 1

Se utiliza la serie de tiempo ARIMA y para determinar sus componentes p, d y q se aplica la prueba de Dicky Fuller de autoregresión y estacionariedad.

**Figura 14**

*Prueba de Dicky Fuller para estacionariedad*

```
result = adfuller(X_train[target].dropna())
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
```

```
ADF Statistic: -3.116909
p-value: 0.025327
```

*Nota*: Resultado de script para evaluar la estacionariedad

El valor de p < 0.05 indica que con 95% de confianza se rechaza la Ho, es decir la serie temporal es no estacionaria lo que significa que será necesario modelar la serie con un enfoque ARIMA.

**Figura 15**

*Resultado de modelo ARIMA con valor p,d,q (3,2,1)*

ARIMA Model Results

| | | | |
|---|---|---|---|
| Dep. Variable: | D2.Venta_zona_3 | No. Observations: | 43 |
| Model: | ARIMA(3, 2, 1) | Log Likelihood | -143.487 |
| Method: | css-mle | S.D. of innovations | 6.417 |
| Date: | Mon, 03 Oct 2022 | AIC | 300.974 |
| Time: | 02:43:16 | BIC | 313.303 |
| Sample: | 2 | HQIC | 305.521 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.4113 | 0.353 | 1.164 | 0.245 | -0.282 | 1.104 |
| Venta_zona_3_lag2_avg3 | -0.0281 | 0.028 | -1.019 | 0.308 | -0.082 | 0.026 |
| ar.L1.D2.Venta_zona_3 | -0.4224 | 0.155 | -2.729 | 0.006 | -0.726 | -0.119 |
| ar.L2.D2.Venta_zona_3 | -0.2348 | 0.178 | -1.319 | 0.187 | -0.584 | 0.114 |
| ar.L3.D2.Venta_zona_3 | -0.0782 | 0.172 | -0.455 | 0.649 | -0.415 | 0.259 |
| ma.L1.D2.Venta_zona_3 | -1.0000 | 0.063 | -15.872 | 0.000 | -1.123 | -0.877 |

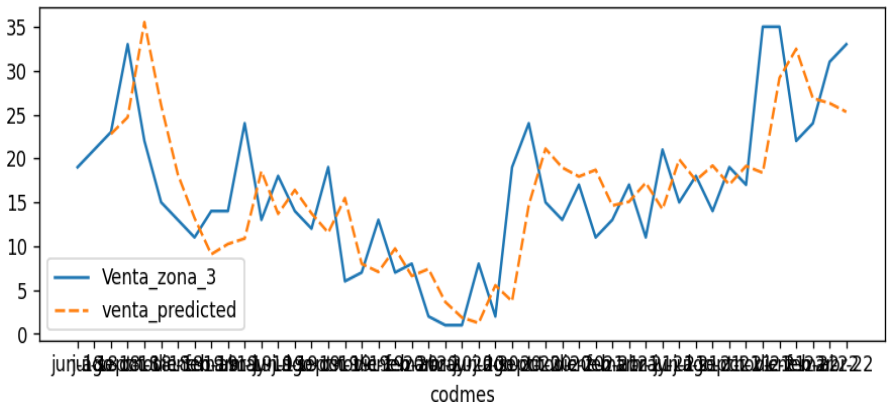
Roots

| | Real | Imaginary | Modulus | Frequency |
|---|---|---|---|---|
| AR.1 | -0.1331 | -2.1579j | 2.1620 | -0.2598 |
| AR.2 | -0.1331 | +2.1579j | 2.1620 | 0.2598 |
| AR.3 | -2.7361 | -0.0000j | 2.7361 | -0.5000 |
| MA.1 | 1.0000 | +0.0000j | 1.0000 | 0.0000 |

*Nota.* Resultado de script para estimar modelo ARIMA

**Figura 16**

*Gráfico de data original versus datos predecidos por el Modelo ARIMA*

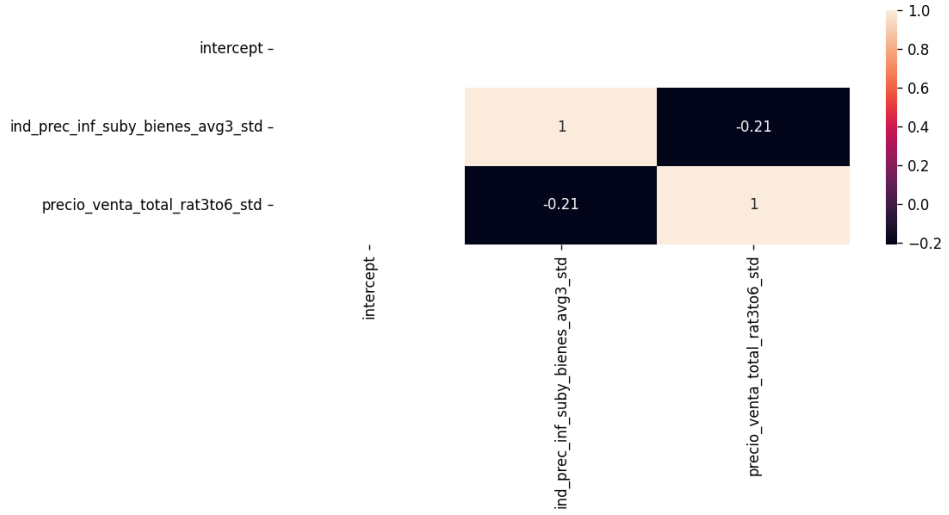


## Etapa 2

Para el modelo de regresión, se eligió por expertis las variables significativas

## Figura 17

*Correlación entre variables y ventas*




## Figura 18

*Resultados de modelo de regresión*

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Venta_zona_3   R-squared:                       0.475
Model:                            OLS   Adj. R-squared:                  0.449
Method:                 Least Squares   F-statistic:                     18.12
Date:                Mon, 03 Oct 2022   Prob (F-statistic):           2.50e-06
Time:                        02:47:43   Log-Likelihood:                -136.59
No. Observations:                  43   AIC:                             279.2
Df Residuals:                      40   BIC:                             284.5
Df Model:                           2
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
intercept                        15.1896      0.918     16.542      0.000      13.334      17.045
ind_prec_inf_suby_bienes_avg3_std     5.2374      0.942      5.559      0.000       3.333       7.142
precio_venta_total_rat3to6_std      -1.0323      0.940     -1.098      0.279      -2.932       0.867
==============================================================================
Omnibus:                        0.101   Durbin-Watson:                   1.498
Prob(Omnibus):                  0.951   Jarque-Bera (JB):                0.138
Skew:                           0.098   Prob(JB):                        0.933
Kurtosis:                       2.803   Cond. No.                         1.24
==============================================================================
```

*Nota.* Resultado de script para estimar modelo de Regresión Múltiple

Las variables cumplen con el criterio de significancia del 95% ($p<0.05$) en su mayoría, sin embargo, la variable precio de venta no, pese a esto se mantendrá la variable dentro de la regresión debido a que es importante mantener este efecto macro en la ecuación. Otro criterio importante que deberán cumplir las variables escogidas será que no se encuentren altamente correlacionadas. El análisis de correlaciones mostrado debajo evidencia que no hay niveles los suficientemente altos como para descartar la independencia de variables.

## Etapa 3

En este paso desarrollaremos un ensamble de medias para ensamblar las predicciones obtenidas por el modelo ARIMA y el modelo de regresión. Esto debido a que los resultados obtenidos por los modelos de ensamble por vectores autorregresivos no obtuvieron resultados esperado a nivel de predicción.

**Tabla 4**

*Resultados de ensamble de zona 3 de Miraflores*

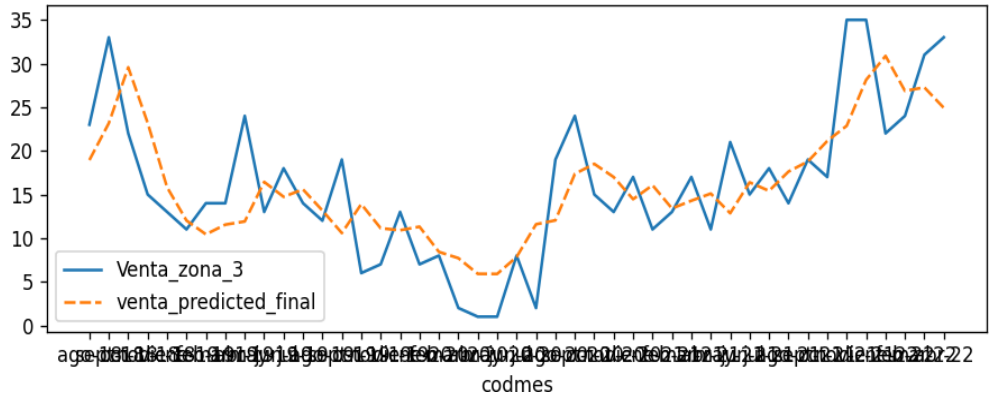
| codmes | Venta_zona_3 | venta_predicted | macro_pred | venta_predicted_final |
|---|---|---|---|---|
| **ago-18** | 23 | 22.784008 | 15.101709 | 18.942858 |
| **sep-18** | 33 | 24.691473 | 21.64755 | 23.169512 |
| **oct-18** | 22 | 35.518131 | 23.615519 | 29.566825 |
| **nov-18** | 15 | 26.064008 | 20.332715 | 23.198362 |
| **dic-18** | 13 | 18.184105 | 13.461249 | 15.822677 |
| **ene-19** | 11 | 13.206798 | 10.825007 | 12.015902 |
| **feb-19** | 14 | 9.074129 | 11.795756 | 10.434943 |
| **mar-19** | 14 | 10.245845 | 12.844848 | 11.545346 |
| **abr-19** | 24 | 10.869423 | 12.95251 | 11.910967 |
| **may-19** | 13 | 18.543232 | 14.333703 | 16.438467 |
| **jun-19** | 18 | 13.690842 | 15.779486 | 14.735164 |
| **jul-19** | 14 | 16.399467 | 14.76305 | 15.581259 |
| **ago-19** | 12 | 13.736599 | 12.581217 | 13.158908 |
| **sep-19** | 19 | 11.541494 | 9.625762 | 10.583628 |
| **oct-19** | 6 | 15.483635 | 12.31068 | 13.897157 |
| **nov-19** | 7 | 7.997548 | 14.267709 | 11.132628 |
| **dic-19** | 13 | 7.070958 | 14.735791 | 10.903374 |
| **ene-20** | 7 | 9.727624 | 12.89158 | 11.309602 |
| **feb-20** | 8 | 6.603906 | 10.241371 | 8.422638 |
| **mar-20** | 2 | 7.408325 | 8.009961 | 7.709143 |
| **abr-20** | 1 | 3.650209 | 8.172444 | 5.911327 |

| | | | | |
|---|---|---|---|---|
| **may-20** | 1 | 1.871414 | 9.930206 | 5.90081 |
| **jun-20** | 8 | 1.244598 | 14.439676 | 7.842137 |
| **jul-20** | 2 | 5.534252 | 17.631289 | 11.58277 |
| **ago-20** | 19 | 3.723152 | 20.359455 | 12.041303 |
| **sep-20** | 24 | 14.684228 | 20.036666 | 17.360447 |
| **oct-20** | 15 | 21.106904 | 15.898517 | 18.50271 |
| **nov-20** | 13 | 18.957431 | 15.027541 | 16.992486 |
| **dic-20** | 17 | 17.931207 | 10.994528 | 14.462868 |
| **ene-21** | 11 | 18.692826 | 13.431659 | 16.062243 |
| **feb-21** | 13 | 14.628739 | 12.173015 | 13.400877 |
| **mar-21** | 17 | 15.069727 | 13.497916 | 14.283821 |
| **abr-21** | 11 | 17.220325 | 12.983888 | 15.102107 |
| **may-21** | 21 | 14.204208 | 11.506386 | 12.855297 |
| **jun-21** | 15 | 19.885521 | 12.891282 | 16.388402 |
| **jul-21** | 18 | 17.50237 | 13.329666 | 15.416018 |
| **ago-21** | 14 | 19.17916 | 16.057403 | 17.618282 |
| **sep-21** | 19 | 17.018014 | 20.540484 | 18.779249 |
| **oct-21** | 17 | 19.129109 | 23.133658 | 21.131383 |
| **nov-21** | 35 | 18.363022 | 27.370057 | 22.86654 |
| **dic-21** | 35 | 29.173773 | 27.082377 | 28.128075 |
| **ene-22** | 22 | 32.477869 | 29.253597 | 30.865733 |
| **feb-22** | 24 | 26.851468 | 26.879208 | 26.865338 |
| **mar-22** | 31 | 26.289792 | 28.232518 | 27.261155 |
| **abr-22** | 33 | 25.300563 | 24.544754 | 24.922658 |

***Nota:*** Contenido de resultado de script para elaboración de estudio.

## Figura 19

*Gráfica de modelo de ensamble*

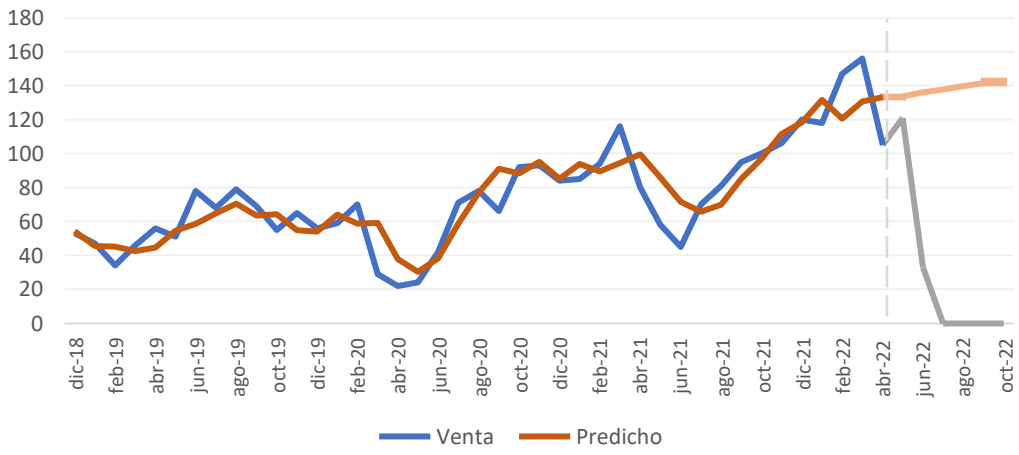


***Miraflores total***

Veamos gráficamente el resultado de los valores predichos para la venta de inmuebles en el distrito de Miraflores.

**Figura 20**

*Ventas del distrito de Miraflores versus los valores predichos*




Con un valor de ajuste de modelo $R^2$ de 79.9% y el valor accuracy de 85% (% de predicciones correctas), podemos evidenciar que el uso del modelo fue el adecuado para las predicciones de ventas de departamentos.

# V. CONCLUSIONES

Con el método de ensamble por medias se obtuvo un valor de predicción del 85% para las ventas estimadas de departamentos en el distrito de Miraflores. Este resultado permitirá que la empresa decida si deberá continuar invirtiendo en mejorar sus sistemas de marketing, canales de ventas, inversión en terrenos, etc.

En la Zona 1 de Miraflores obtuvo un valor de predicción del 59.1%, en la Zona 2 de Miraflores del 79.5% y la Zona 3 del 75.0%.
El valor de R2 para predecir las ventas de departamentos del distrito de Miraflores, alcanzó un valor de 79.9%, indicando que el modelo es robusto.

Las variables creadas con ratio promedio de 3 a 6 meses alcanzaron ingresar en el modelo, por lo que realizar nuevamente el estudio a partir de estas variables podrían mejorar los resultados.

# VI. RECOMENDACIONES

Se recomienda usar acumulaciones trimestrales por sobre las estimaciones mensuales además de agrupaciones por distrito por sobre las desagregaciones por zona del distrito. Esto con la finalidad de mantener los niveles de precisión de estimación lo más robustos y confiables posible. Además de seguir probando con otras variables para la predicción de las ventas al ser las variables exógenas fluctuantes, y seguir investigando mercados que hayan elaborado modelos para la estimación de las futuras ventas de inmuebles.

# VII. REFERENCIAS BIBLIOGRAFICAS


BBC News Mundo (8 de setiembre de 2020). Cómo se explica el insólito "boom" inmobiliario en medio de la peor crisis económica de las últimas décadas. https://www.bbc.com/mundo/noticias-54035630

Gonzáles, M. P. (2009). Técnicas de predicción económica. Universidad del País Vasco. 145pp.

Hernández, J. (2008). Análisis de series temporales económicas I (2° ed.) ESIC Editorial. 86pp.

Jaume, M. J. y Catalá, R.M. (2001). Estadística Informática: Casos y ejemplos con el SPSS. Universidad de Alicante. 336pp.

Libro sin autor: (“Estadística II 4ta edición, unidad 7”, s.f.)

Nieves, V. y Becedas, M. (28 de octubre de 2022). La caída del sector inmobiliario amenaza con sepultar la economía global por tres vías diferentes. El Economista. https://www.eleconomista.es/vivienda-inmobiliario/noticias/12009030/10/22/La-caida-del-sector-inmobiliario-amenaza-con-sepultar-la-economia-global-por-tres-vias-diferentes.html

Patel, H. (2021). What is Feature Engineering - Importance, Tools and Techniques for Machine Learning. Towards Data Science.

Postgrado UTP (16 de octubre de 2021). Sector inmobiliario en Perú: efectos y recuperación post-pandemia. https://www.postgradoutp.edu.pe/blog/a/sector-inmobiliario-en-peru-efectos-y-recuperacion-post-pandemia/

PricewaterhouseCoopers (s.f.). Manual de Forecasting. Instituto Aragones de Fomento.

Ricarey, R. (2021). Métodos de ensamblado en Machine Learning. [Trabajo de fin de Máster titulado]. Universidade de Santiago de Compostela.

Ríos, G. y Hurtado, C. (2008). Series de tiempo. Universidad de Chile. 52pp.

Vásquez, M. (2017). Panorama inmobiliario internacional. Real Estate México. https://realestatemarket.com.mx/articulos/mercadoinmobiliario/21574-panorama-inmobiliario-internacional