PROYECTO INTEGRADOR

Predicción del índice de precios de la vivienda nueva en Medellín y el Área Metropolitana

Presentado por:
David Betancur Londoño
Diego Andrés Jaramillo Zapata
Susana Londoño Muñoz
Catalina Piedrahita Jaramillo
Diego Andrés Valderrama Laverde

Universidad Eafit
Maestría en Ciencias de los Datos y Analítica
Medellín
2020-Semestre II

Contenido

Resumen

En el presente trabajo se aplican técnicas estadísticas y de aprendizaje de máquina aprendidas durante los diferentes cursos desarrollados en el semestre. Con estas técnicas se pretende predecir el Índice de Precios de la Vivienda Nueva (IPVN) en Medellín y el área metropolitana, y entender su comportamiento con respecto a otras variables económicas e indicadores asociados al sector de la construcción. Se proyectan los valores del 2020 y 2021, y se evalúa el desempeño de cada método para dicha tarea.

Para esto, se buscaron fuentes públicas de datos económicos como el DANE, Camacol y el Banco de la República de donde se recolectaron datos trimestrales, de carácter agregado, asociados con el sector de la construcción y la actividad productiva, tanto a nivel local como nacional. Con los datos reunidos, se organizó un conjunto de datos de 42 variables y 61 observaciones.

El desarrollo del ejercicio se basa en la metodología CRISP-DM. Comenzando con la clarificación del problema, donde se expone cómo, para el caso puntual de Colombia y Medellín y el área metropolitana, el IPVN puede reflejar en cierta medida la interacción entre la oferta y la demanda de vivienda nueva, y cómo su evolución se ha venido presentando en una senda creciente en los últimos años, sumado a que su pronóstico podría ser de interés en el marco de aumentar el acceso efectivo a vivienda. Seguido de la descripción y recolección de los datos, resultados de la exploración e ingeniería de características, que permiten contar con un conjunto de datos con menos variables, pero de mayor significancia estadística y práctica.

Posteriormente, se implementan los modelos a partir de dos enfoques, el multivariado y el univariante. En el primer caso, se entiende el IPVN y su relación con otras variables, utilizando las técnicas de regresión lineal múltiple y redes neuronales. En el caso del acercamiento univariante, se aplica el análisis de series de tiempo, con la técnica de Box-Jenkins, donde se modela el pronóstico del IPVN a partir de su comportamiento histórico, sin considerar la incidencia de otras variables; además, en el enfoque univariante se aplica redes neuronal LSTM, un tipo de red recurrente, que se fundamenta en el trabajo con datos secuenciales.

Finalmente, con el fin de determinar el modelo de mejor pronóstico, se utiliza como criterio el error porcentual medio absoluto, por sus siglas en inglés, MAPE. Donde las técnicas univariadas presentan una mayor precisión. Tanto en la modelación multivariada como univariante, las redes neuronales presentaron mejores resultados frente a las técnicas estadísticas tradicionales, sin embargo, requieren de un mayor esfuerzo computacional y mayor experimentación. Con los modelos de mejor precisión se obtiene el pronóstico del año 2020 y 2021.

Palabras claves: índice de precio de la vivienda nueva; pronóstico; regresión lineal múltiple; redes neuronales; series de tiempo.

Clarificación del problema

En el caso colombiano, el acceso a la vivienda propia, por parte de las familias, es un proyecto de mediano y largo plazo, que es de gran relevancia en la construcción de patrimonio familiar y en el mejoramiento de condiciones de vida. En este proyecto, normalmente familiar, interactúan dos fuerzas:

- La oferta de viviendas, nuevas y usadas, según la zona de interés. La cual depende a su vez de la disponibilidad de suelos, costos de los insumos, cumplimientos normativos, entre otros.
- La demanda, dada principalmente por la disponibilidad de recursos de las familias o de inversores para comprar las viviendas. Acá desempeña un rol clave tanto los ahorros que puedan tener las personas, como también, el acceso efectivo a crédito. Este último, está fuertemente influenciado por el cumplimiento de las condiciones de acceso y el valor de la tasa de interés. Destacando que, en los últimos años, la concurrencia de subsidios estatales también ha desempeñado un papel preponderante, al permitir un aumento de la disponibilidad de recursos.

La interacción de ambas fuerzas, que se conoce tradicionalmente como mercado, es lo que genera la formación de precios. Esta variable es de suma importancia, puesto que se convierte en la mayoría de los casos, en el factor decisivo para el acceso efectivo de las familias a una vivienda digna o deseada.

En los últimos años, en las principales ciudades y áreas metropolitanas del país, entre ellas Medellín y el área metropolitana, se ha presentado que, el precio promedio de la vivienda nueva, medido a través del índice de precios de la vivienda nueva - IPVN que estima trimestralmente el Departamento Administrativo Nacional de Estadística - DANE, ha mostrado un crecimiento continuo, inclusive por encima de la variación del índice de precios al consumidor (inflación). Lo cual, si bien, no ha detenido la construcción, y más importante, el acceso a vivienda nueva, plantea preguntas relacionadas sobre su comportamiento futuro y las variables de oferta y demanda que más influyen en su comportamiento.

Inclusive, entender cómo evolucionará el IPVN, desempeñaría un insumo clave en términos de política económica y social, con la cual se podrían diseñar instrumentos que faciliten el acceso, especialmente, de la población de menores ingresos, además, como un posibilitador para dinamizar el sector de la construcción, de importancia en la dinamización productiva de los territorios. Asimismo, podría ofrecer nueva información, en términos de rentabilidad de inversión futura, tanto para compradores como para constructores.

Objetivos

General

Predecir el índice de precios de la vivienda nueva (IPVN) en Medellín y el área metropolitana, empleando modelos estadísticos y técnicas de aprendizaje de máquina.

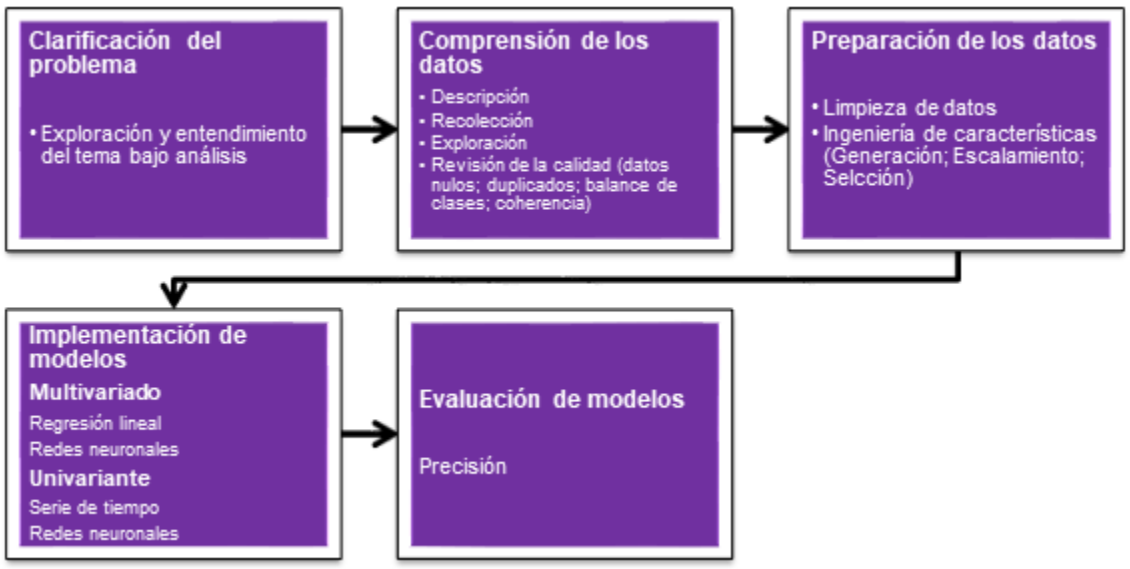
Específicos

- Encontrar variables con mayor influencia en el IPVN.
- Identificar modelos para realizar la predicción del IPVN.
- Medir la capacidad de predicción de modelos estadísticos y de aprendizaje de máquina.

Metodología

Para el cumplimiento de los objetivos trazados, se utiliza la metodología CRISP-DM, de acuerdo con las siguientes etapas:

Metodología implementada. Adaptada de la metodología general CRISP-DM



Comprensión de los datos

Descripción

Los datos usados para este trabajo provienen del Departamento Administrativo Nacional de Estadística -DANE-. Los datos específicos del sector de la construcción, tomados del DANE, se encuentran agregados en un informe de la Cámara colombiana de la construcción -Camacol, llamado *Colombia construcción en cifras*. El dataset construido contiene las siguientes 42 variables macroeconómicas y del sector de la construcción:

Variable	Descripción	Unidades
d_TD_MDE_Porc	Tasa de desempleo de Medellín	Porcentaje
d_IPVN_MDE_Num	Índice de Precio de Vivienda Nueva Medellín Valor Total	Número
d_TO_MDE_Porc	Tasa de ocupación de Medellín	Porcentaje
d_PT_MDE_Num	Población total de Medellín	Millones
d_PET_MDE_Num_Millon	Población en edad de trabajar en Medellín	Millones
d_PEA_MDE_Nun_Millon	Población económicamente activa en Medellín	Millones
d_IPC_MDE_Num	Índice de precio al consumidor	Porcentaje
d_IPC_NAL_Porc	Índice de precio al consumidor	Porcentaje
d_TOSC_Nal_Porc	Tasa de ocupación del sector de la construcción nacional	Porcentaje
d_TDSC_NAL_Porc	Tasa de desocupación del sector de la construcción nacional	Porcentaje
d_UVAPCSL_Total_NAL_Num	Unidades de vivienda aprobadas para construcción según licencias	Número
d_UIVIS_MDEVA_Num	Unidades iniciadas de VIS Valle de Aburrá	Número
d_UINOVIS_MDEVA_Num	Unidades iniciadas de NO VIS Valle de Aburrá	Número
d_UIVISNOVIS_MDEVA_Num	Unidades iniciadas VIS y NO VIS Valle de Aburrá	Número
d_PIB_VPCB2015_NAL_Num_Mil_millon	PIB Valor a precio corriente base 2015 en miles de millones (Datos corregidos de efectos estacionales y de calendario)	Miles de millones
d_PIB_VPKB2015_NAL_Num_Mil_millon	PIB Valor a precio constante base 2015 en miles de millones (Datos corregidos de efectos estacionales y de calendario)	Miles de millones
d_PIB_VPCB2015_SC_Num_Mil_Millon	PIB sector de la construcción valor a precio corriente base 2015 en miles de millones	Miles de millones
d_PIB_VPKB2015_SC_Num_Mil_Millon	PIB sector de la construcción valor a precio constante base 2015 en miles de millones	Miles de millones
d_TIPPBR_NAL_Porc	Tasa de interés promedio ponderado (Promedio trimestral de la tasa de interés de colocación) del Banco de la República %1	Porcentaje
d_TIPPST_NAL_Porc	Tasa de interés promedio ponderado (Promedio trimestral de la tasa de interés de colocación) sin tesorería % 3	Porcentaje
d_OC_MDE_MT2	Obras culminadas en el área metropolitana de Medellín, metros cuadrados.	Metros cuadrados
d_UVAPCSL_Total_ANT_Num	Unidades de vivienda aprobadas para construcción según licencias	Número
d_UVAPCSL_VIS_ANT_Num	Unidades de vivienda aprobadas para construcción según licencias, Vivienda de interés social, Antioquia. Número de unidades	Número
d_UVAPCSL_NoVIS_Ant_Num	Unidades de vivienda aprobadas para construcción según licencias, NO vivienda de interés social, Antioquia. Número de unidades	Número
d_ALCSD_ANT_VIS_mt2	ANTIOQUIA: Área (m2) licenciada para construcción según destino, vivienda de interés social	Metros cuadrados
d_ALCSD_NoVis_ANT_mt2	ANTIOQUIA: Área (m2) licenciada para construcción según destino, NO vivienda de interés social	Metros cuadrados
d_ICCV_NAL_Num	Índice de costos de construcción de vivienda (ICCV) -	Número

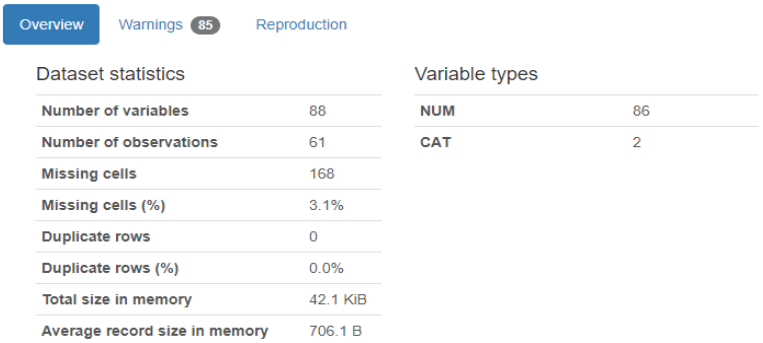
d_ICCV_Mat_NAL_Num	Índice de costos de construcción de vivienda (ICCV) - Material	Número
d_ICCV_MO_NAL_Num	Índice de costos de construcción de vivienda (ICCV) - Mano de Obra	Número
d_ICCV_MaqEq_NAL_Num	Índice de costos de construcción de vivienda (ICCV) - Maquinaria y Equipo	Número
d_ICCV_MDE_Num	Índice de Costos de Construcción de Vivienda (ICCV)	Número
d_IPPTotal_NAL_Num	Índice de Precios al Productor (IPP) - Total	Número
d_IPPMat_NAL_Num	Índice de Precios al Productor (IPP) - Materiales de Construcción	Número
d_IPVUNom_NAL_Num	Índice de precios de vivienda usada Nominal	Número
d_IPVUReal_NAL_Num	Índice de precios de vivienda usada Real	Número
d_AFC_NAL_Millon	Cuentas de Ahorro para el Fomento de la Construcción - AFC, saldo y número	Millones
d_AFC_NAL_Numero	Cuentas de Ahorro para el Fomento de la Construcción - AFC, saldo y número	Número
d_CAPVis_NAL_Millon	Cuentas de Ahorro Programado - CAP para VIS, saldo y número	Millones
d_CAPVis_NAL_Num	Cuentas de Ahorro Programado - CAP para VIS, saldo y número	Número
d_PDCG_NAL_Ton	Producción de Cemento Gris en Toneladas - Nacional	Toneladas
d_DDCG_NAL_Ton	Despacho de Cemento Gris en Toneladas - Nacional	Toneladas
d_DDCG_ANT_Ton	Despacho de Cemento Gris en Toneladas - Antioquia	Toneladas

Para cada una de las variables se hizo el cálculo de la variación trimestral anual, de forma que el conjunto de datos final contiene 84 variables.

Exploración

Exceptuando las etiquetas de periodo y trimestre, todos los datos son numéricos. No se identificaron registros nulos o duplicados. Es de esperar este resultado dado que la información es oficial, reportada principalmente por el Dane, y dadas sus políticas de entrega de información se espera que esta cuente con altos niveles de calidad.

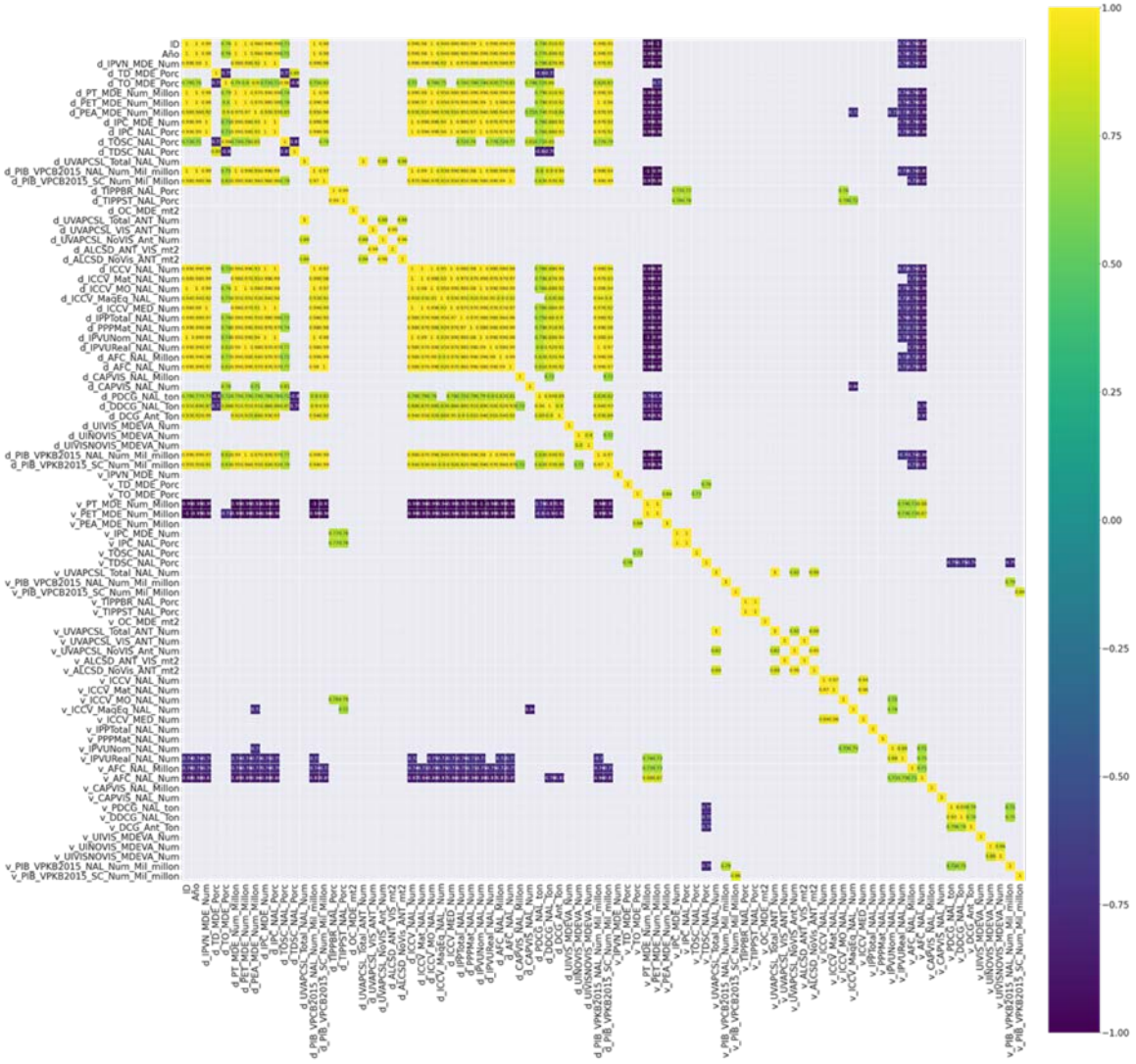
El resumen de la exploración utilizando *pandas profiling* es el siguiente:



Los datos ausentes corresponden a los cuatro trimestres de 2005 en las variaciones. Esto se debe a que las variaciones se calculan con respecto al año anterior y no se tienen datos de 2004. Además, se identificaron ocho variables con datos atípicos.

Matriz de correlación

Al conjunto de datos se le hizo una matriz de correlación que permitiera ver gráficamente las relaciones de la variable respuesta con otras variables, y de las demás variables entre ellas.



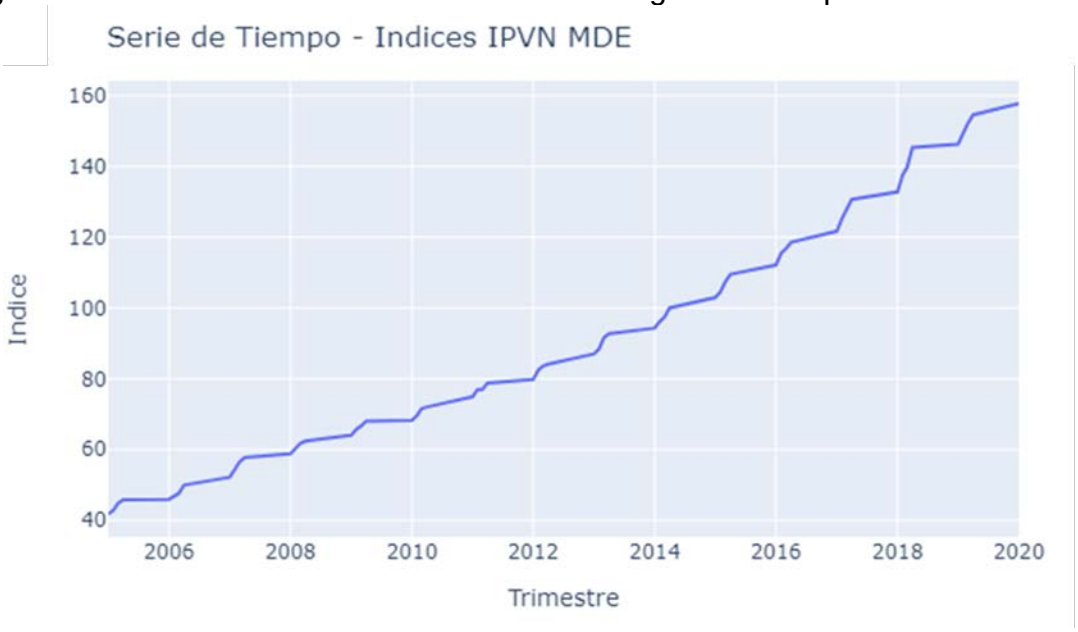
En el análisis de esta matriz se puede ver como la variable respuesta [d_IPVN_MDE_Num] está correlacionada con otros datos y variaciones. Existe poca o nula correlación de la variación del índice con otras variables.

Exploración de la variable respuesta: IPVN

El siguiente es el resultado de las principales estadísticas descriptivas de la variable respuesta:

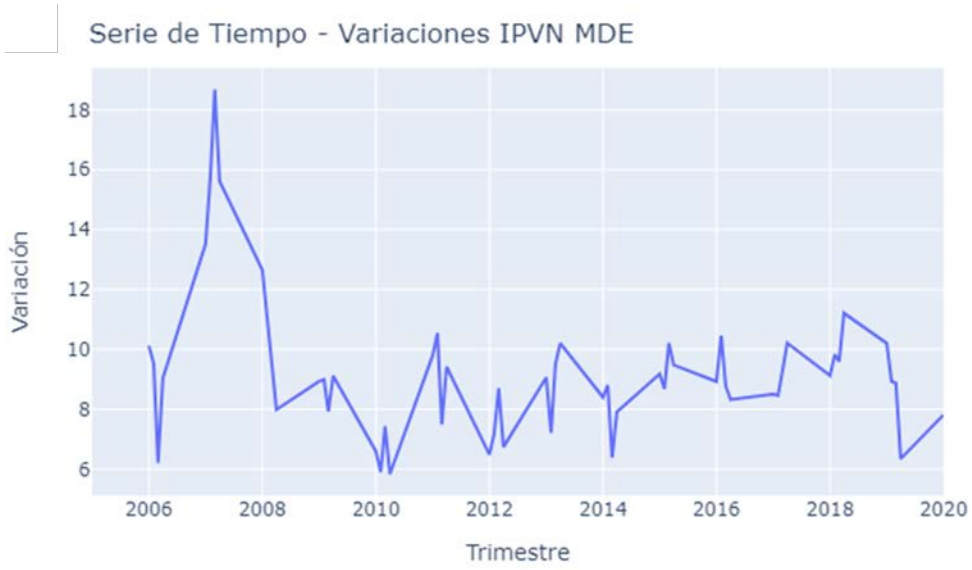
	ID	Año	d_IPVN_MDE_Num	v_IPVN_MDE_Num
count	61.000000	61.000000	61.000000	57.000000
mean	31.000000	2012.131148	89.652459	9.233520
std	17.752934	4.440253	33.447189	2.366486
min	1.000000	2005.000000	41.720000	5.835661
25%	16.000000	2008.000000	62.350000	7.924161
50%	31.000000	2012.000000	83.600000	8.988391
75%	46.000000	2016.000000	115.480000	9.828343
max	61.000000	2020.000000	157.770000	18.680395

Al graficar la evolución del índice se observa el siguiente comportamiento:



Este índice ha tenido un comportamiento creciente con periodos de aceleración.

La variación trimestral anual del índice ha tenido ascensos y descensos, pero desde el año 2008 no se observan cambios bruscos en el comportamiento. El promedio de la variación ha sido del 9.28%



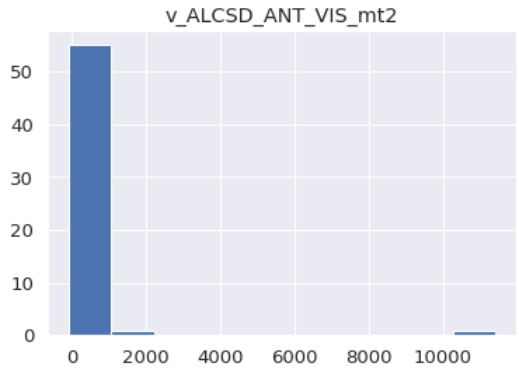
Preparación de los datos

Ingeniería de características

Se hace ingeniería de características sobre las variables a usar en los modelos con enfoque multivariado.

Identificación de datos atípicos

Se identificaron ocho características en las cuales la variación trimestral anual es atípica al compararse con los demás registros. Estas características fueron eliminadas. Por ejemplo: v_ALCSD_ANT_VIS_Mt2: Área (m2) licenciada para construcción según destino, vivienda de interés social:



Escalamiento de los datos

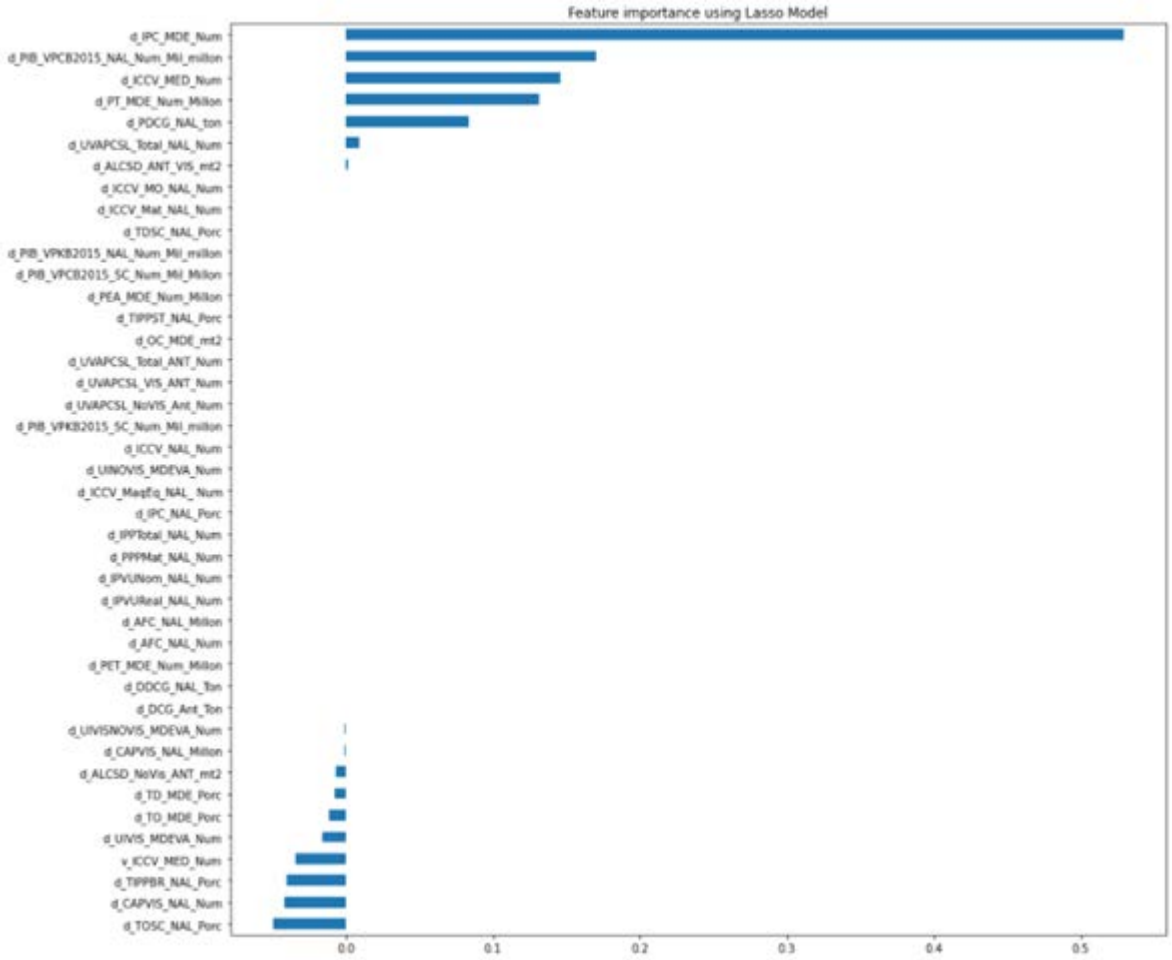
La base de datos contaba con múltiples variables macroeconómicas del sector construcción que tenían diferentes unidades, algunas en valor de tasas, como la tasa de desempleo u ocupación y otras en grandes valores monetarios como el Producto Interno Producto, por lo que primero se decidió llevar a cabo una

normalización, restando la media y dividiendo por su desviación estándar, esto para poder realizar comparación entre las variables y evitar que algunas tuvieran un mayor peso en los modelos.

Selección de características

Para realizar una primera selección de características se utilizó un método integrado, la regresión de Lasso con penalización, esto significa que se ha incluido un término regulador Alpha, con el que se busca dar una sanción a los coeficientes para lograr un modelo más simple; esto provoca que algunos coeficientes relacionados a características que se consideran poco relevantes para estimar la variable respuesta terminen siendo cero. Cuando se realiza este proceso parte del reto está en definir el Alpha óptimo, por ello se decidió usar la función LassoCV de la librería Sklearn que ayuda encontrar el Alpha óptimo para la regularización.

En total, con este método se escogieron 17 variables. A continuación, se muestra la gráfica que indica el peso de las variables seleccionadas.



Con el primer resultado obtenido en la selección de características, se procedió a estimar una regresión lineal, cuyo resultado fue el siguiente:

```
Call:
lm(formula = d_IPVN_MDE_Num ~ d_IPC_MDE_Num + d_PIB_VPCB2015_NAL_Num_Mil_millon +
    d_ICCV_MED_Num + d_PT_MDE_Num_Millon + d_PDCG_NAL_ton + d_UVAPCSL_Total_NAL_Num +
    d_ALCSD_ANT_VIS_mt2 + d_UIVISNOVIS_MDEVA_Num + d_CAPVIS_NAL_Millon +
    d_ALCSD_NoVis_ANT_mt2 + d_TD_MDE_Porc + d_TO_MDE_Porc + d_UIVIS_MDEVA_Num +
    v_ICCV_MED_Num + d_TIPPBR_NAL_Porc + d_CAPVIS_NAL_Num + d_TOSC_NAL_Porc,
    data = dataN2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.041045 -0.014944 -0.000722  0.010403  0.058613

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -0.015476   0.007355  -2.104 0.043861 *
d_IPC_MDE_Num                   0.378677   0.079164   4.783 4.29e-05 ***
d_PIB_VPCB2015_NAL_Num_Mil_millon  0.301106   0.165771   1.816 0.079315 .
d_ICCV_MED_Num                  0.334393   0.110887   3.016 0.005182 **
d_PT_MDE_Num_Millon            -0.073447   0.168850  -0.435 0.666685
d_PDCG_NAL_ton                  0.047512   0.021151   2.246 0.032198 *
d_UVAPCSL_Total_NAL_Num         0.032352   0.019536   1.656 0.108142
d_ALCSD_ANT_VIS_mt2            -0.012985   0.009742  -1.333 0.192594
d_UIVISNOVIS_MDEVA_Num         -0.000441   0.007878  -0.056 0.955728
d_CAPVIS_NAL_Millon             0.021981   0.010256   2.143 0.040316 *
d_ALCSD_NoVis_ANT_mt2          -0.027417   0.016607  -1.651 0.109193
d_TD_MDE_Porc                  -0.014062   0.011688  -1.203 0.238312
d_TO_MDE_Porc                  -0.009704   0.021725  -0.447 0.658322
d_UIVIS_MDEVA_Num              -0.005197   0.008091  -0.642 0.525565
v_ICCV_MED_Num                 -0.037412   0.007912  -4.729 5.01e-05 ***
d_TIPPBR_NAL_Porc              -0.031091   0.007664  -4.057 0.000326 ***
d_CAPVIS_NAL_Num               -0.031894   0.013094  -2.436 0.021015 *
d_TOSC_NAL_Porc                -0.007938   0.022342  -0.355 0.724844
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02611 on 30 degrees of freedom
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9988
F-statistic:  2252 on 17 and 30 DF,  p-value: < 2.2e-16
```

Como puede apreciarse en el resultado de la regresión, muchas de las variables no son significativas, ya que el p-valor es mayor a 0.05. Dado entonces que algunas variables no son significativas y el ajuste de la regresión es alto con un R^2 de 0.99, se decidió filtrar más características; para esta segundo filtro se utilizó la función step() del paquete R, con esta función se busca lograr el mejor modelo basado en el criterio de información de Akaike (AIC), el cual busca un equilibrio entre la bondad de ajuste y la complejidad del modelo. Como resultado de este proceso se obtiene un modelo con 12 variables:

```
Call:
lm(formula = d_IPVN_MDE_Num ~ d_IPC_MDE_Num + d_PIB_VPCB2015_NAL_Num_Mil_millon +
    d_ICCV_MED_Num + d_PDCG_NAL_ton + d_UVAPCSL_Total_NAL_Num +
    d_ALCSD_ANT_VIS_mt2 + d_CAPVIS_NAL_Millon + d_ALCSD_NoVis_ANT_mt2 +
    d_TD_MDE_Porc + v_ICCV_MED_Num + d_TIPPBR_NAL_Porc + d_CAPVIS_NAL_Num,
    data = dataN2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.044797 -0.015460  0.000165  0.011190  0.067908

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -0.021380   0.005112  -4.182 0.000184 ***
d_IPC_MDE_Num                   0.353956   0.065667   5.390 4.92e-06 ***
d_PIB_VPCB2015_NAL_Num_Mil_millon  0.192786   0.069138   2.788 0.008506 **
d_ICCV_MED_Num                  0.381126   0.087686   4.346 0.000113 ***
d_PDCG_NAL_ton                  0.040283   0.018378   2.192 0.035131 *
d_UVAPCSL_Total_NAL_Num         0.036624   0.017976   2.037 0.049229 *
d_ALCSD_ANT_VIS_mt2            -0.016070   0.008808  -1.825 0.076621 .
d_CAPVIS_NAL_Millon             0.027560   0.007793   3.536 0.001165 **
d_ALCSD_NoVis_ANT_mt2          -0.029767   0.015472  -1.924 0.062528 .
d_TD_MDE_Porc                  -0.012236   0.009581  -1.277 0.209982
v_ICCV_MED_Num                 -0.033955   0.005816  -5.839 1.26e-06 ***
d_TIPPBR_NAL_Porc              -0.031257   0.006799  -4.597 5.38e-05 ***
d_CAPVIS_NAL_Num               -0.041305   0.010038  -4.115 0.000224 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02501 on 35 degrees of freedom
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9989
F-statistic:  3478 on 12 and 35 DF,  p-value: < 2.2e-16
```

Dado que aún se obtuvieron variables que no son significativas y el ajuste del modelo era muy alto con un R^2 ajustado de 0.99, se decidió eliminar de forma iterativa las variables con un p-valor mayor a 0.05, esto para lograr una mayor parsimonia en el modelo. Se utilizó entonces una eliminación secuencial hacia atrás, se partió de la última regresión, se eliminó la variable con el mayor valor p, se estimó de nuevo y, así sucesivamente, hasta obtener un modelo con todas las variables significativas.

A continuación, el resultado de este proceso. Un modelo con 8 variables, todas ellas significativas bajo el criterio del valor p.

```
Call:
lm(formula = d_IPVN_MDE_Num ~ d_IPC_MDE_Num + d_PIB_VPCB2015_NAL_Num_Mil_millon +
    d_ICCV_MED_Num + d_PDCG_NAL_ton + d_CAPVIS_NAL_Millon + v_ICCV_MED_Num +
    d_TIPPBR_NAL_Porc + d_CAPVIS_NAL_Num, data = dataN2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.056035 -0.016898 -0.000751  0.011252  0.065715

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -0.020043   0.004868  -4.117 0.000193 ***
d_IPC_MDE_Num                   0.418312   0.062532   6.690 5.73e-08 ***
d_PIB_VPCB2015_NAL_Num_Mil_millon  0.212772   0.070162   3.033 0.004296 **
d_ICCV_MED_Num                  0.293930   0.082039   3.583 0.000932 ***
d_PDCG_NAL_ton                  0.055273   0.013492   4.097 0.000205 ***
d_CAPVIS_NAL_Millon             0.025385   0.007309   3.473 0.001274 **
v_ICCV_MED_Num                 -0.034473   0.005345  -6.449 1.23e-07 ***
d_TIPPBR_NAL_Porc              -0.032088   0.006949  -4.617 4.15e-05 ***
d_CAPVIS_NAL_Num               -0.045930   0.009584  -4.793 2.40e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02607 on 39 degrees of freedom
Multiple R-squared:  0.999,     Adjusted R-squared:  0.9988
F-statistic:  4801 on 8 and 39 DF,  p-value: < 2.2e-16
```

Las ocho características seleccionadas son las que se consideran como base para comenzar la estimación de los modelos.

Implementación de modelos

A continuación, se presentan los modelos implementados, se utilizaron 2 enfoques, modelos multivariados, en los cuales se explica la variable respuesta a partir de otras variables económicas y del sector, y modelos univariantes donde se explica la variable a partir de su comportamiento histórico.

Multivariado

Regresión lineal múltiple

Con el ajuste de una regresión lineal múltiple se busca crear un modelo que permita identificar el conjunto de variables que mejor describen la variable respuesta que, para este caso, es el Índice de Precios de la Vivienda Nueva (IPVN) de Medellín y el área metropolitana.

Tomando como punto de partida las ocho variables identificadas en la ingeniería de características, se realizó la siguiente matriz de correlación:

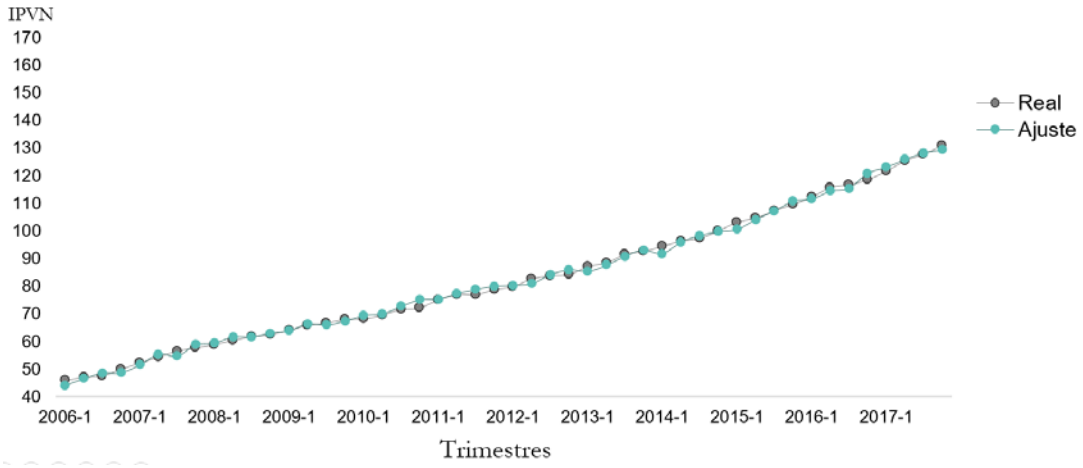
Puede apreciarse que existe una alta correlación entre algunos de los predictores, por lo que se procedió a retirar 3 de ellos del modelo. A continuación, la versión final del modelo de regresión estimado para el IPVN con 5 variables.

```
Call:
lm(formula = d_IPVN_MDE_Num ~ d_IPC_MDE_Num + d_PDCG_NAL_ton +
    v_ICCV_MED_Num + d_TIPPBR_NAL_Porc + d_CAPVIS_NAL_Num, data = dataN2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.096679 -0.026502 -0.009114  0.022474  0.083055

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.019098   0.007264  -2.629 0.011907 *
d_IPC_MDE_Num      0.886701   0.013943  63.596  < 2e-16 ***
d_PDCG_NAL_ton     0.121539   0.010105  12.027 3.44e-15 ***
v_ICCV_MED_Num    -0.028010   0.007330  -3.821 0.000432 ***
d_TIPPBR_NAL_Porc -0.044660   0.008611  -5.186 5.80e-06 ***
d_CAPVIS_NAL_Num  -0.054548   0.011063  -4.931 1.33e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0408 on 42 degrees of freedom
Multiple R-squared:  0.9973,    Adjusted R-squared:  0.997
F-statistic:  3130 on 5 and 42 DF,  p-value: < 2.2e-16
```

Finalmente, las variables del modelo son:

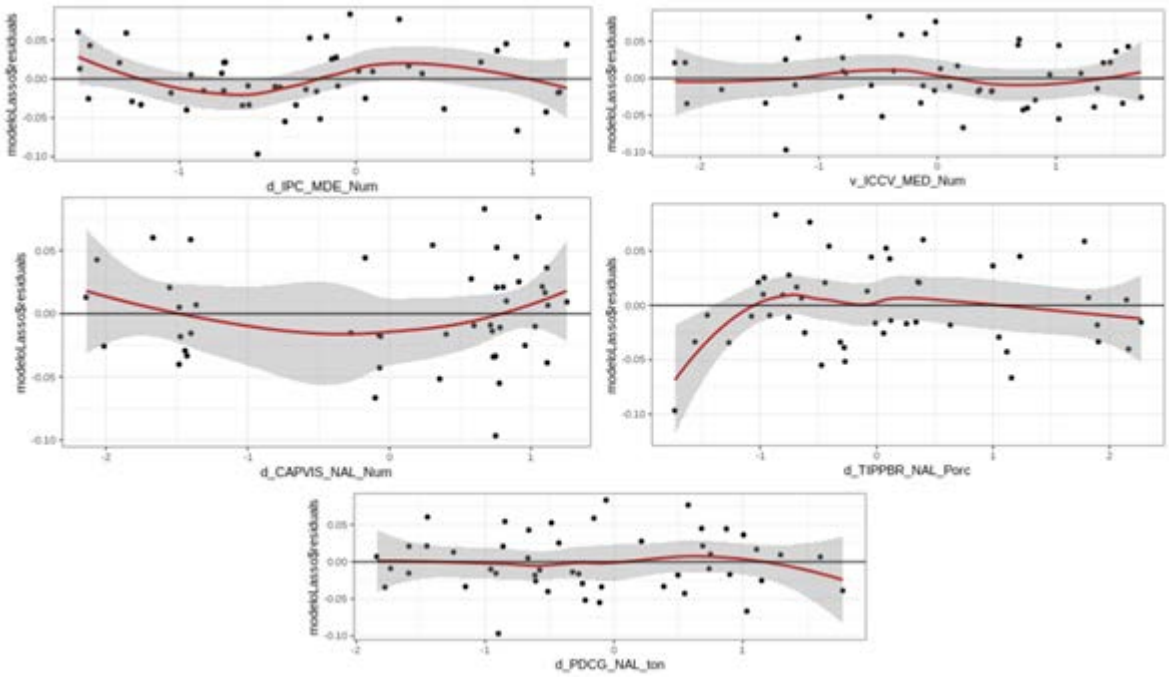
- *d_IPC_MDE_Num*: Índice de precio al consumidor de Medellín y el área metropolitana.
- *d_TIPPBR_NAL_Porc*: Tasa de interés promedio ponderado (Promedio trimestral de la tasa de interés de colocación) del Banco de la República
- *v_ICCV_MDE_Num*: Índice de Costos de Construcción de Vivienda (ICCV) de Medellín y el área metropolitana.
- *d_CAPVis_NAL_Num*: Cuentas de Ahorro Programado - CAP para VIS, número.
- *d_PDCG_NAL_Ton*: Producción de Cemento Gris en Toneladas – Nacional

A continuación, se muestra el ajuste de la regresión desde el 2016-01 hasta el 2017-04



Validación de supuestos

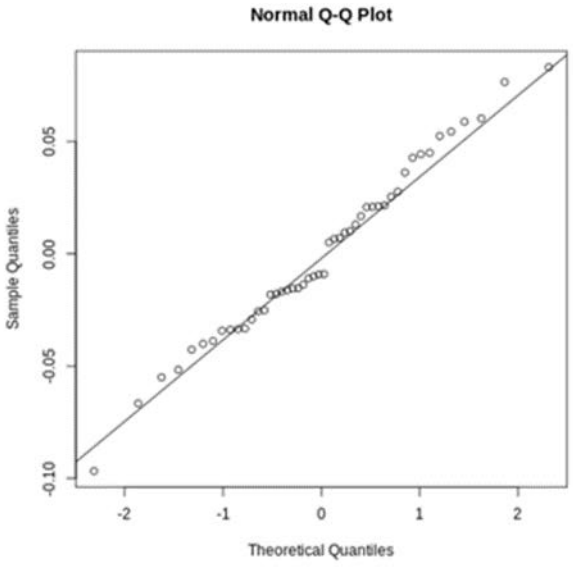
Primero se procedió a revisar los diagramas de dispersión entre cada uno de los predictores y los residuos del modelo, en los cuales no se deben apreciar patrones, sino un comportamiento de los residuos aleatorio en torno a 0 con una variabilidad constante a lo largo del eje X. Este análisis se realiza para validar la relación lineal entre los predictores y la variable respuesta.



Se concluyó que se cumple la linealidad para todos los predictores, al no observarse patrones en los residuales versus cada predictor

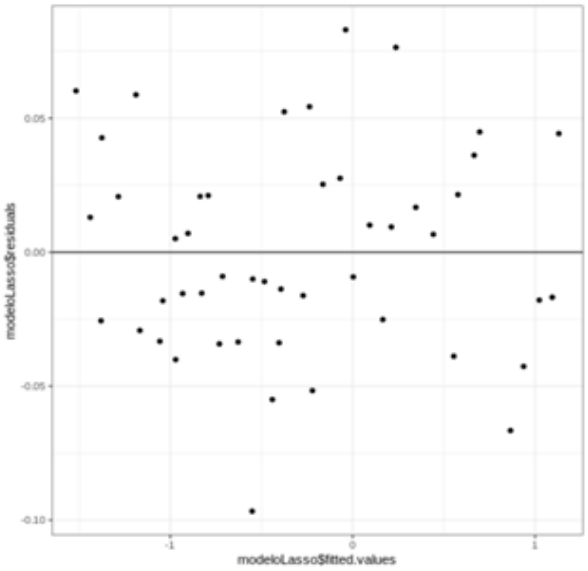
Ahora se procede a evaluar la hipótesis de normalidad de los residuos, utilizando para ello la prueba de Shapiro-Wilk:

```
        Shapiro-Wilk normality test

data:  modeloLasso$residuals
W = 0.98608, p-value = 0.8338
```



Con el resultado anterior se concluye que no hay evidencia suficiente para negar el supuesto de normalidad de los residuales.

Posteriormente, se procedió a evaluar la homocedasticidad de los residuos, es decir, que tengan una variabilidad constante; para ello se realizó la gráfica de los residuales versus los valores ajustados, en esta gráfica busca validarse que no se presenten patrones y que los puntos tengan un comportamiento aleatorio en torno a cero a lo largo del eje X.



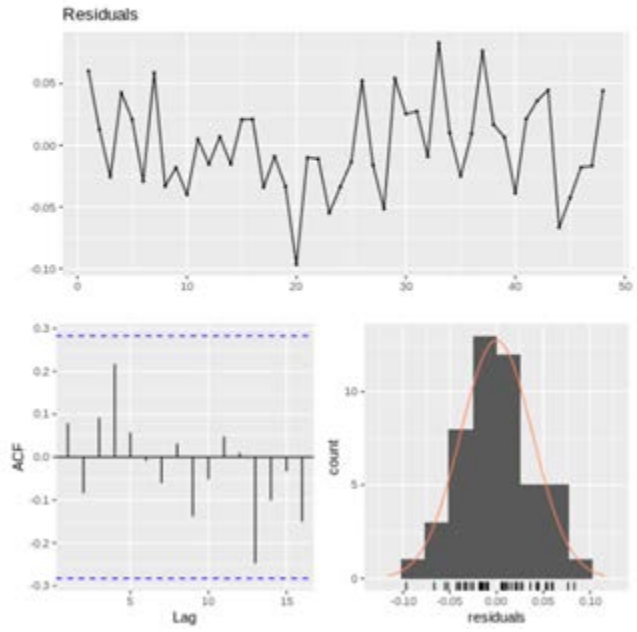
Se concluye con la gráfica anterior que no hay evidencia de falta de homocedasticidad.

También se realizó la prueba de Breusch-Pagan para revisar la homocedasticidad, obteniendo como resultado un p-valor mayor a 0.05, lo cual confirma la conclusión anterior.

```
        studentized Breusch-Pagan test

data:  modeloLasso
BP = 1.2981, df = 5, p-value = 0.9351
```

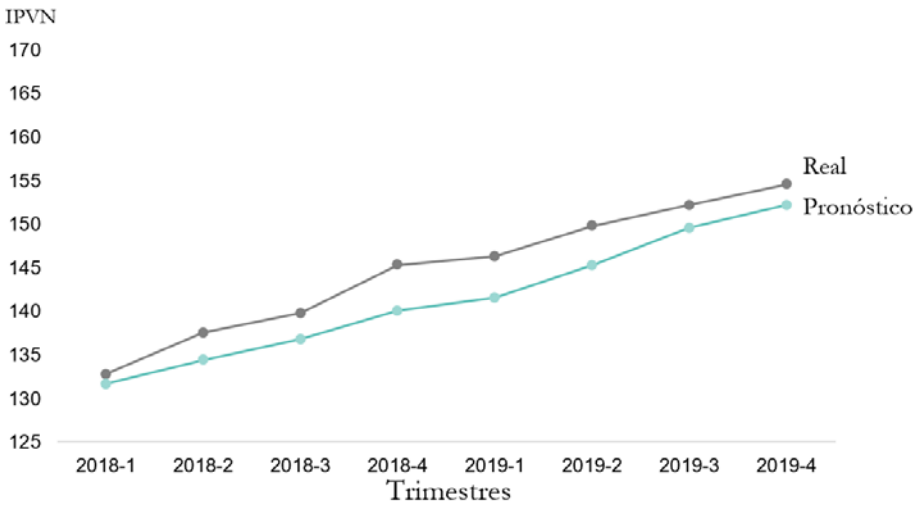
Se procede a revisar la autocorrelación de los residuos. Concluyendo con la prueba de Breusch-Godfrey que no hay evidencia de autocorrelación

```
        Breusch-Godfrey test for serial correlation of order up to 15

data:  Residuals
LM test = 14.079, df = 15, p-value = 0.5195
```



Evaluación

Se seleccionaron los últimos 8 datos, como la muestra de pruebas, y sobre los restantes se estimó el modelo anterior, en el cual se validaron los supuestos sobre los residuales de normalidad, homocedasticidad y autocorrelación. Con base en esto, se hizo el pronóstico, obteniendo los siguientes resultados con un error porcentual medio absoluto de 2.62%



Interpretación de resultados

Puede apreciarse en los resultados de la regresión lineal que la variable del modelo con mayor peso para la variable respuesta del IPVN es *d_IPC_MDE_Num*: índice de precios al consumidor de la ciudad, su relación es directamente proporcional, el aumento en el IPC significa un aumento en el IPVN, esta relación tiene sentido ya que ambas variables miden una relación de precios, por lo que un aumento de los precios en la ciudad puede tener un efecto en el precio de la vivienda.

La segunda variable con mayor peso en el modelo es *d_PDCG_NAL_Ton:* producción de cemento gris en toneladas, la variable que tiene el modelo hace referencia a las toneladas nacionales, y puede apreciarse que tiene un impacto significativo para el IPVN de Medellín y el Área Metropolitana, esto puede explicarse dado que es la segunda ciudad más importante del país. La relación entre esta variable y la variable respuesta es positiva, lo que nos indica que un crecimiento en las toneladas producidas significa aumentos en el índice, esta relación tiene sentido ya que quizás un aumento de las toneladas significa un crecimiento del sector de la construcción, es decir, tal vez nuevas unidades de vivienda que pueden afectar por ende el valor del índice de precios de vivienda nueva.

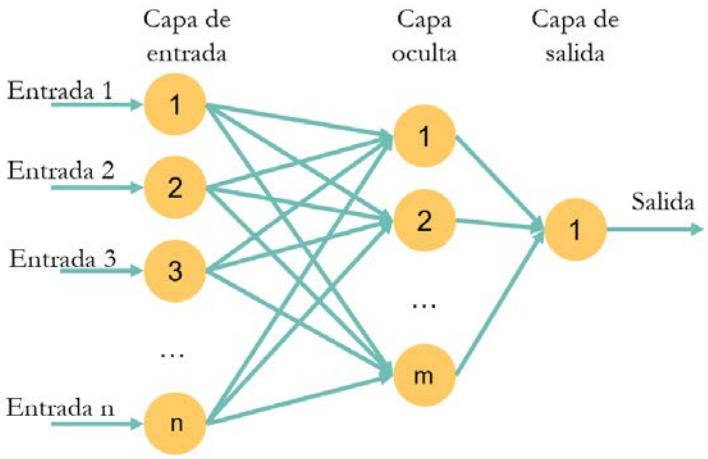
Las demás variables como *v_ICCV_MDE_Num:* la variación trimestral anual del índice de costos de construcción de vivienda, *d_CAPVis_NAL_Num*: la cantidad de cuentas de ahorro programado de Vivienda de Interés Social (VIS) y *d_TIPPBR_NAL_Porc*: la Tasa de interés promedio ponderado trimestral de colocación del Banco de la República tienen una relación negativa con la variable respuesta, es decir, que un aumento en ellas significa una disminución en el índice.

Al convertir los datos normalizados a sus correspondientes valores, tenemos una descripción del impacto de las unidades de la siguiente manera para las dos variables más importantes.

- El incremento en 1 punto del IPC de la ciudad significa un incremento de 2.09 puntos para el IPVN, si todo lo demás permanece constante
- La producción de cemento gris tiende a caer para los primeros trimestres del año y aumentar para el último, en promedio tiene una variación de 40.000 toneladas por mes, por eso usamos esta cifra como base para el análisis. Un incremento de 40.000 en las toneladas producidas significa un incremento de 0.5 puntos para el IPVN, si todo lo demás permanece constante.

Redes neuronales

También exploramos la alternativa de emplear Redes Neuronales Artificiales (RNA), las cuales son un modelo inspirado en el funcionamiento del cerebro humano. Está formado por un conjunto de nodos conocidos como neuronas artificiales que están conectadas y transmiten señales entre sí.



Grafica con base a: Que son las redes neuronales y sus funciones (Enlace)

Su objetivo es aprender modificándose automáticamente a sí misma de forma que pueda llegar a realizar tareas complejas. Aplicada a la ciencia de datos, las redes neuronales pueden realizar predicciones gracias a su capacidad generalizadora por el hecho de aprender a partir de ejemplos.

Particularmente para nuestro caso de estudio, la entrada de la red neuronal es resultado de la ingeniería de características, donde se obtuvo un conjunto de datos con pocas variables de forma que:

- Se gana eficiencia computacional
- Minimiza la experimentación en la combinatoria de parámetros
- Hace más comparables los métodos

En cuanto a la parametrización de la red neuronal, por la cantidad de registros del set de datos, es necesario tener cuidado de no crear un modelo demasiado complejo, lo que podría llevar a sobre ajustar los resultados. Para ello se adoptó

una arquitectura basada en dos capas densas, la primera con 1024 y la segunda con 512 neuronas, ambas utilizando una función de activación RELU (Rectified Linear Unit). Se utilizó también una capa densa con una función RELU como capa de salida.

Para saber si el modelo está aprendiendo correctamente, se utilizó una función de pérdida de error cuadrático medio y para reportar su desempeño la métrica de Error Porcentual Absoluto Medio (MAPE). Al utilizar el método de resumen de Keras, podemos ver que tenemos un total de 534.529 parámetros, lo cual es aceptable para el conjunto de datos.

```
from keras.models import Sequential
from keras.layers import Dense
model = Sequential()
model.add(Dense(1024, input_shape=(8, ), activation='relu', name='dense_1'))
model.add(Dense(512, activation='relu', name='dense_2'))
model.add(Dense(1, activation='relu', name='dense_output'))
model.compile(optimizer='adam', loss='mse', metrics=['mape'])
model.summary()
```

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 1024)              9216
_________________________________________________________________
dense_2 (Dense)              (None, 512)               524800
_________________________________________________________________
dense_output (Dense)         (None, 1)                 513
=================================================================
Total params: 534,529
Trainable params: 534,529
Non-trainable params: 0
```

Con el fin de que el método fuese comparable con el de regresión lineal, los datos de pruebas fueron los últimos 8 trimestres comprendidos entre el 2018-01 al 2019-04

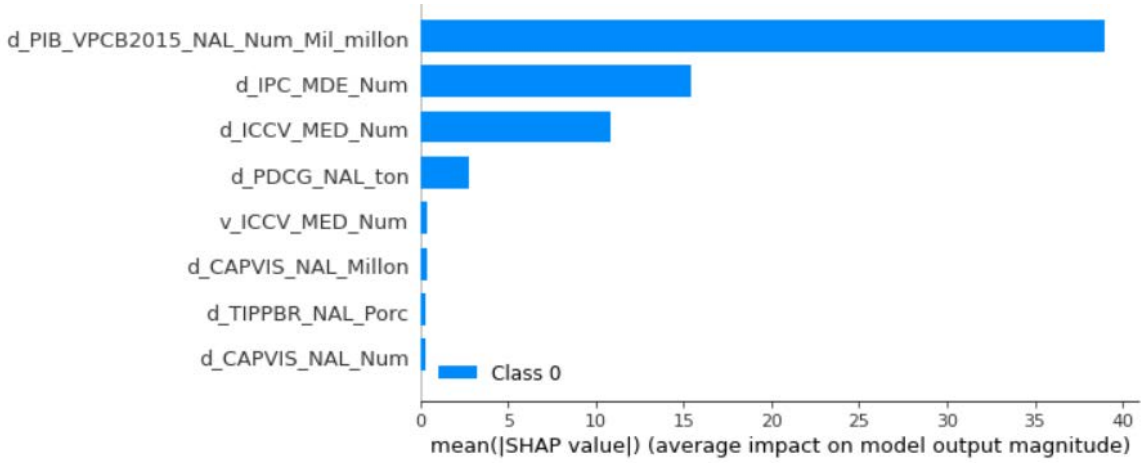
Una vez se entrenó el modelo, se evaluaron los resultados para identificar qué tan confiables son sus predicciones. Como anteriormente se mencionó, esta evaluación se realizó mediante el error cuadrático medio, el cual nos entrega un resultado de 18.72 considerado aceptable.

El Error Porcentual Absoluto Medio (MAPE), que es un indicador del desempeño del pronóstico que mide el tamaño del error (absoluto) en términos porcentuales entrega un resultado de 2.2836 considerado también bastante bueno.

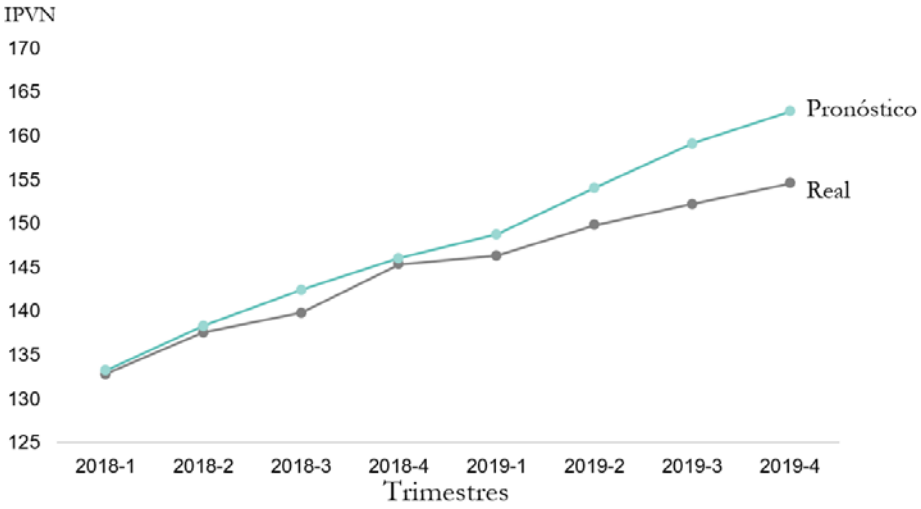
Si bien las redes neuronales son llamadas cajas negras debido a que no se conoce con certeza cómo el algoritmo aprende y toma decisiones; sí podemos conocer

cuáles son las variables más relevantes y que tomaron un mayor protagonismo a la hora de realizar el entrenamiento y su posterior predicción.

Esto quiere decir que para la red neuronal el PIB Nacional es la variable con mayor peso en la predicción de la variable objetivo.



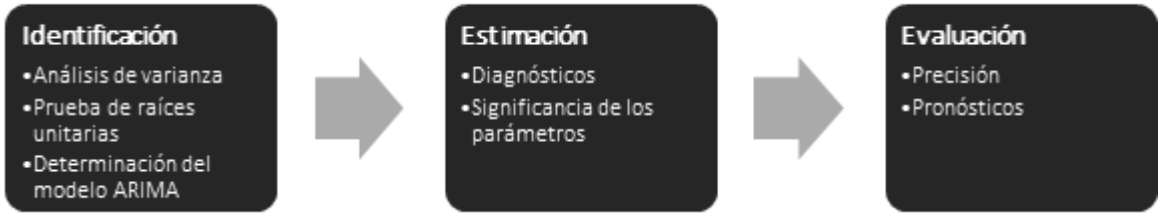
Para finalizar el análisis, a continuación, se muestra gráficamente como el modelo se ajusta y podría predecir adecuadamente la variable objetivo, desplegando el valor real vs el predicho entre los trimestres 2018-01 y 2019-04.



Univariante

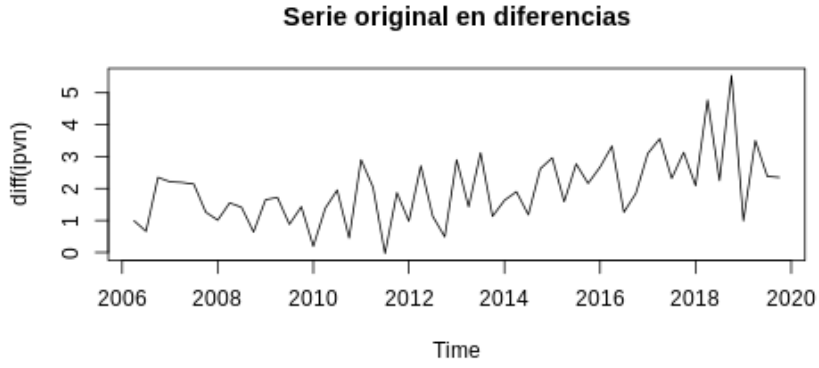
Serie de tiempo

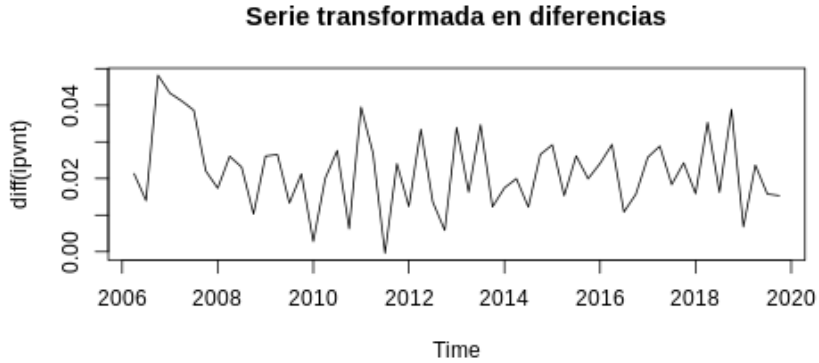
Para la implementación de esta técnica se siguieron, de manera general, los siguientes pasos, a partir de Castaño (2020):



Identificación

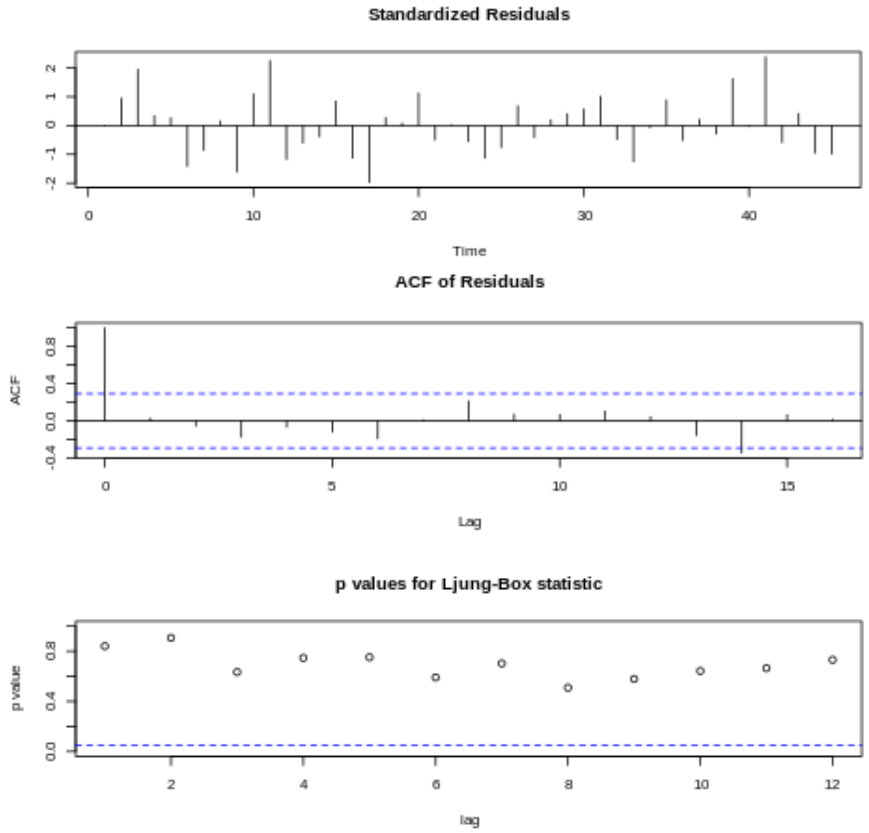
En esta etapa se tomó como punto de partida el análisis gráfico, donde se evidencia que es una serie no estacionaria, que requiere estabilización de varianza y aplicación de diferencias.



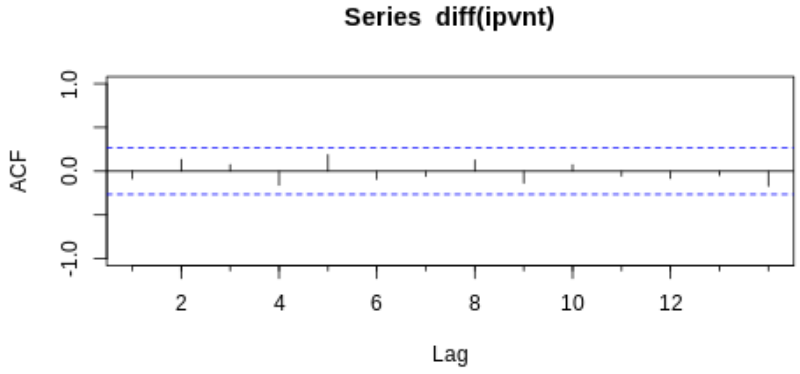


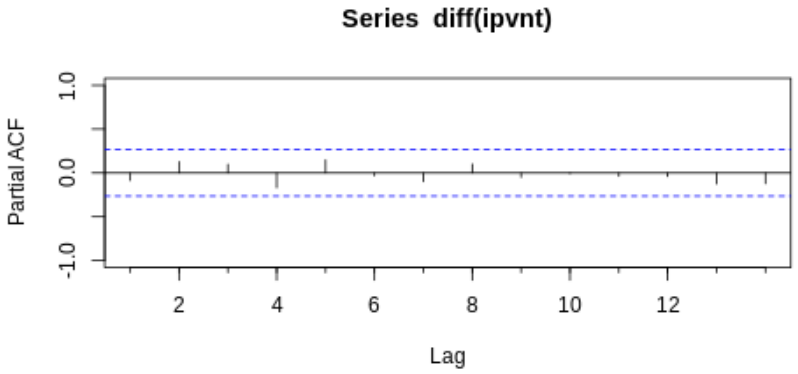
Para la estabilización de varianza, se aplicó la técnica de Box - Cox, obteniéndose un lambda cercano a cero, cuyo resultado implica la transformación vía logaritmo. Dado que es una serie con tendencia creciente se aplicó la primera diferencia y se validó que tuviese al menos una raíz unitaria. Asimismo, se aplicó prueba para

determinar si requería la aplicación de más diferencias, encontrando en las pruebas que no era necesario. En todos los casos, el modelo cumple que los residuales son ruido blanco.



Para la determinación del modelo ARIMA se calcularon y graficaron, inicialmente, los correlogramas de la serie transformada, evidenciándose que la serie no tiene componente autoregresiva o de media móvil. Lo que llevó a plantear un modelo ARIMA (0,1,0) con deriva.





Para validar esto, se utilizaron las siguientes técnicas:

- Implementación del algoritmo auto.arima (R), que arrojó también el mismo modelo ARIMA (0,1,0).

```
# selección "automática" del modelo usando librería foreast
auto.arima(ipvnt,max.p=10, max.q=10, ic=c("aic"))

Series: ipvnt
ARIMA(0,1,0) with drift

Coefficients:
       drift
      0.0221
s.e.  0.0014

sigma^2 estimated as 0.0001117:  log likelihood=172.77
AIC=-341.54   AICc=-341.31   BIC=-337.52
```

- Estimación iterativa de los criterios de información de AIC y BIC para diferentes p (5) y q (5), organizándose de menor a mayor. En este caso, se observa como el modelo ARIMA (0,1,0) es el que obtuvo el menor BIC y el segundo menor AIC.

```
cbind(AIC[order(Aic),], "     ", BIC[order(Bic),])

   p d q       Aic "     " p d q       Bic
15 2 1 2 -341.7414        0 1 0 -337.5246
1  0 1 0 -341.5392        1 1 0 -333.9256
16 2 1 3 -340.5481        0 1 1 -333.8393
7  1 1 0 -339.9476        0 1 2 -331.2959
```

- Uso de la función de autocorrelación extendida muestral (EACF), con la que se validó que el modelo a utilizar debe ser un ARIMA (0,1,0).

```
eacf(diff(ipvnt))
```

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 o o o o o o o o o o o  o  o  o
1 x o o o o o o o o o o  o  o  o
2 x x o o o o o o o o o  o  o  o
3 x o o o o o o o o o o  o  o  o
4 x o x o o o o o o o o  o  o  o
5 o o o o o o o o o o o  o  o  o
6 x o o o o o o o o o o  o  o  o
7 x o x o o o o o o o o  o  o  o
```

Finalmente, se concluye que el modelo a estimar es un ARIMA (0,1,0) con deriva.

Estimación

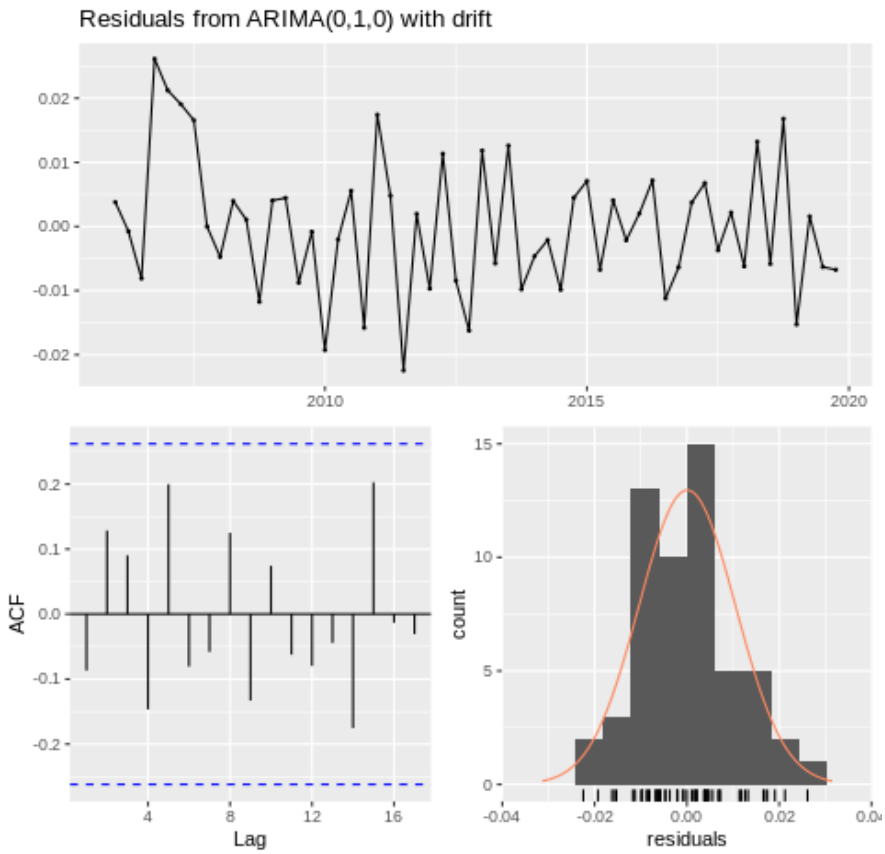
Se estimó el modelo ARIMA (0,1,0) con deriva, obteniéndose los siguientes resultados:

```
mod1_CSS_ML=Arima(ipvn, c(0,1,0), include.drift=TRUE,lambda=0, method = c("CSS-ML"))
summary(mod1_CSS_ML)
```

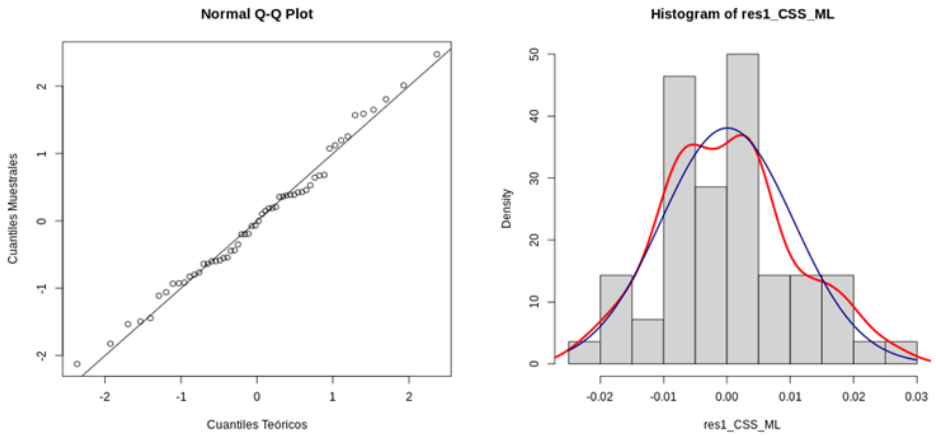
```
Series: ipvn
ARIMA(0,1,0) with drift
Box Cox transformation: lambda= 0

Coefficients:
       drift
      0.0221
s.e.  0.0014

sigma^2 estimated as 0.0001117:  log likelihood=172.77
AIC=-341.54   AICc=-341.31   BIC=-337.52

Training set error measures:
                     ME      RMSE       MAE        MPE      MAPE       MASE
Training set -0.04115317 0.9076744 0.7343985 0.00141472 0.8327413 0.09256302
                    ACF1
Training set -0.3230722
```

Posteriormente, se realizaron los siguientes diagnósticos:

- Los residuales sean ruido blanco. Evidenciándose el cumplimiento de este supuesto, tal como se comprueba a continuación:

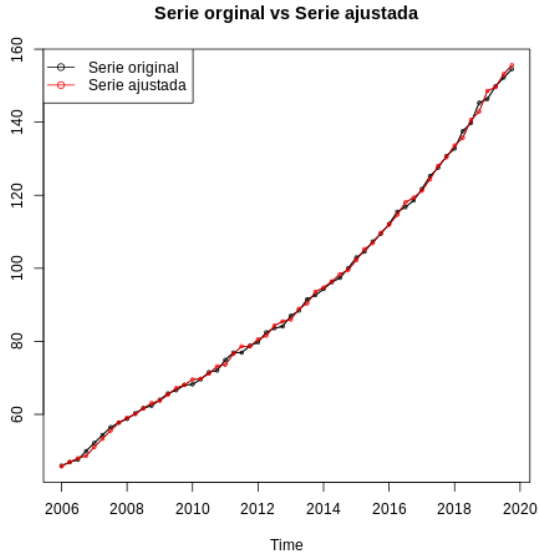


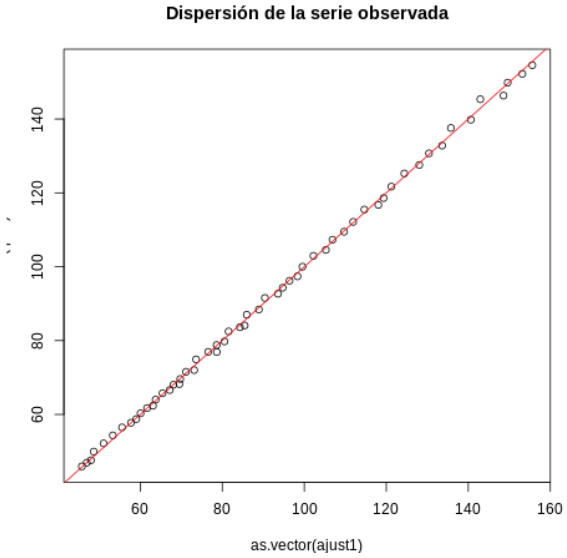
- Los residuales se distribuyen como una normal. Se aplicó la prueba de Shapiro-Wilk, aceptándose la hipótesis nula de que los residuales se distribuyen bajo normalidad y se validó gráficamente:



Asimismo, se validó la significancia del coeficiente estimado, que obtuvo un valor p menor a 0.05.

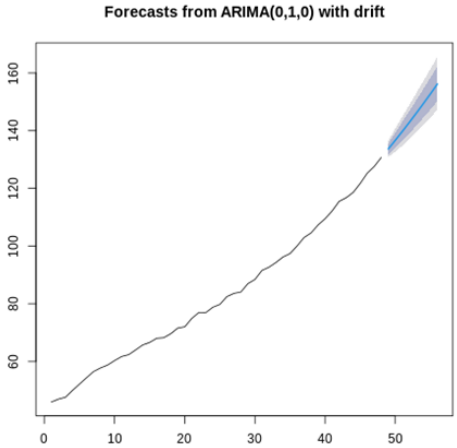
Finalmente, se presenta de manera gráfica el ajuste, observándose que la serie ajustada sigue un comportamiento muy similar a la serie original:

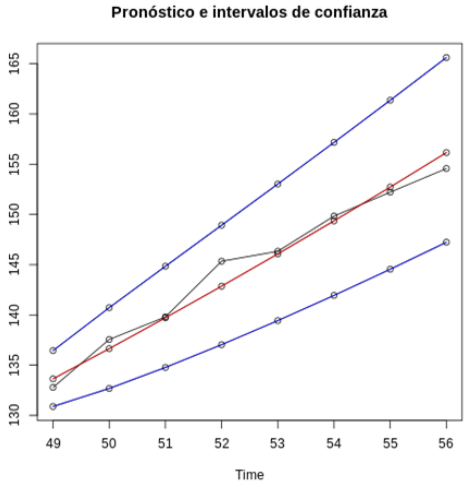




Evaluación

Se seleccionaron los últimos 8 datos como la muestra de prueba, y con los otros datos se estimó el modelo anterior, además, de que se validó el cumplimiento de los supuestos sobre los residuales (ruido blanco y normalidad). Con base en esto, se hizo el pronóstico, cuyos resultados fueron los siguientes:

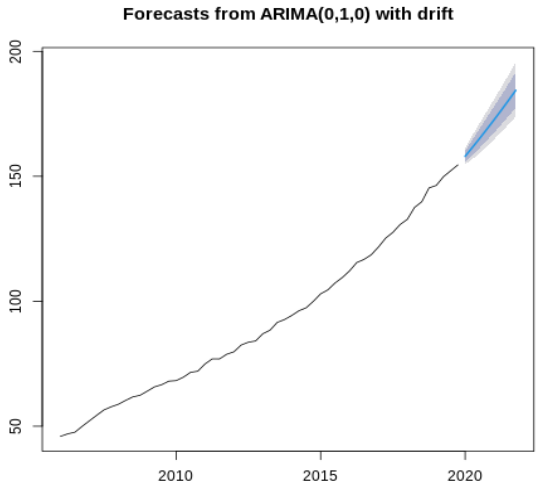




El error porcentual medio absoluto obtenido fue del 0,617%.

Predicción

Finalmente, tomando todo el conjunto de datos, se hizo la predicción hasta el cuarto trimestre de 2021, y se calcularon los intervalos de confianza.



	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2020 Q1	158.0177	155.8919	160.1726	154.7781	161.3252
2020 Q2	161.5424	158.4774	164.6666	156.8786	166.3448
2020 Q3	165.1456	161.3164	169.0658	159.3254	171.1785
2020 Q4	168.8293	164.3171	173.4653	161.9776	175.9707
2021 Q1	172.5950	167.4460	177.9024	164.7828	180.7777
2021 Q2	176.4448	170.6868	182.3971	167.7152	185.6288
2021 Q3	180.3805	174.0308	186.9618	170.7605	190.5424
2021 Q4	184.4039	177.4730	191.6056	173.9100	195.5311

Se evidencia, de acuerdo con este pronóstico, que el IPVN continuará con su tendencia creciente, en consonancia con su evolución histórica.

Redes neuronales

Para este caso, se utilizó una red neuronal recurrente o *Recurrent Neural Networks* (RNN) en inglés, que se usan para analizar datos de series temporales considerando la dimensión de tiempo, basadas en el uso de información secuencial. Este tipo de redes se basan en bucles que llevan a que la salida de la red o de una parte de ella, en un momento dado, sirva como entrada de la propia red en el siguiente momento (Torres, 2019).

Una categoría de estas redes, se denomina *redes de larga memoria a corto plazo* por sus siglas LSTM y es una técnica de aprendizaje profundo. Fueron desarrolladas por Hochreiter and Schmidhuber (1997) y tienen la capacidad de aprender dependencias a largo plazo. Se desarrollaron para hacer frente al problema del gradiente de desaparición que se puede encontrar al entrenar a los RNN tradicionales.

Estimación

Para la implementación de esta técnica, se utilizó la librería Keras de Python, con los siguientes pasos:

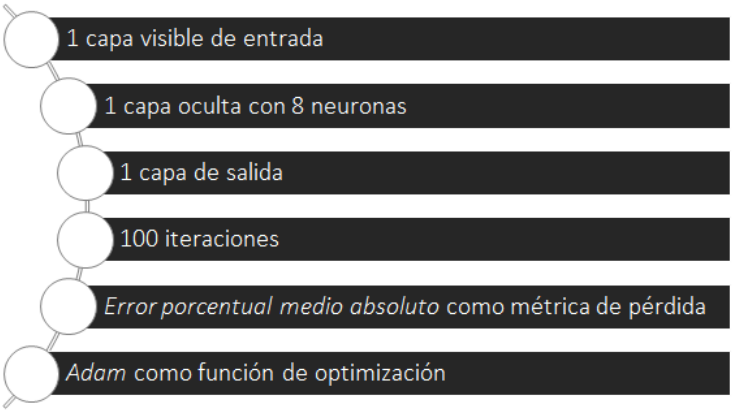
- Se normalizó la variable respuesta - IPVN, dado que este tipo de técnicas son altamente sensibles a la escala. De esta manera, se obtuvieron valores de cero a uno.
- Los datos se dividieron de la siguiente manera: un 80% para entrenamiento y un 20% para prueba, con el fin de hacerlo comparable con los anteriores modelos.
- Se organizaron los datos de la siguiente manera:

Estructura de los datos de la red neuronal		
Muestras (Trimestres)	Características (Valor del índice)	Paso del tiempo

Se seleccionó un *look back* de 3. Seguidamente, se presenta una muestra de la estructura de datos utilizada en la red neuronal:

	0	1	2
0	0.000000	0.008853	0.014755
1	0.008853	0.014755	0.035769
2	0.014755	0.035769	0.055531
3	0.035769	0.055531	0.075114
4	0.055531	0.075114	0.094250
5	0.075114	0.094250	0.105517
6	0.094250	0.105517	0.114549
7	0.105517	0.114549	0.128409
8	0.114549	0.128409	0.141018
9	0.128409	0.141018	0.146741

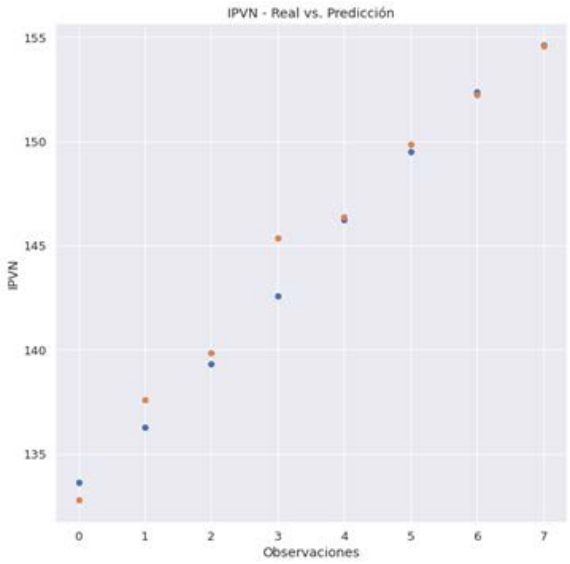
- Se creó la red neuronal con los siguientes parámetros:

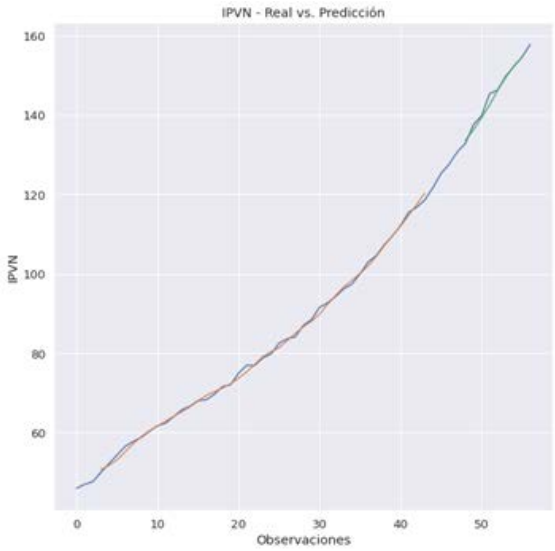


Estos parámetros se obtienen después de explorar algunas combinaciones de los mismos, encontrándose que estos permiten una caída rápida de la métrica de pérdida, sumado a que tienen una buena eficiencia computacional y una alta precisión.

Evaluación

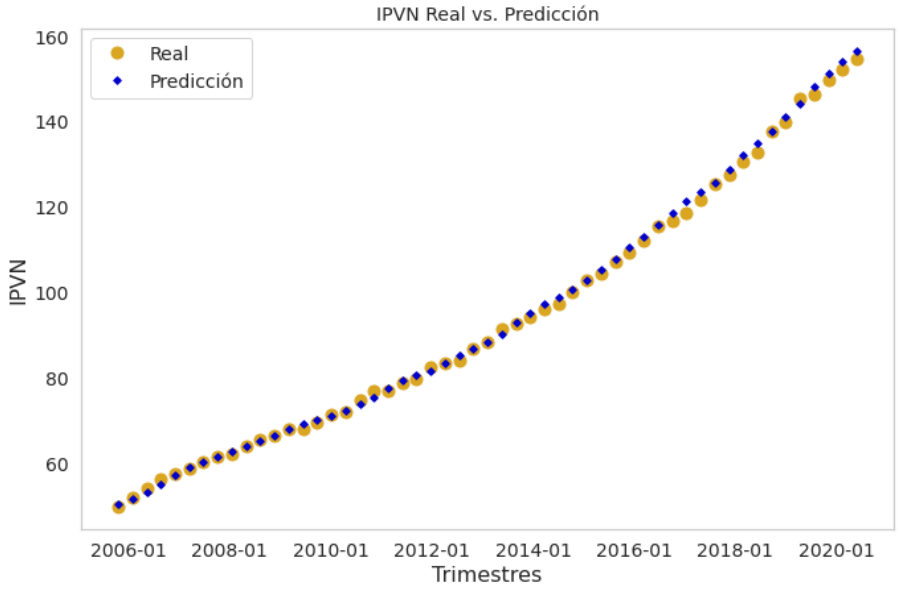
Se obtuvo un error porcentual medio absoluto de 0.531%. y se constató gráficamente un alto ajuste y una alta capacidad predictora por parte del modelo:





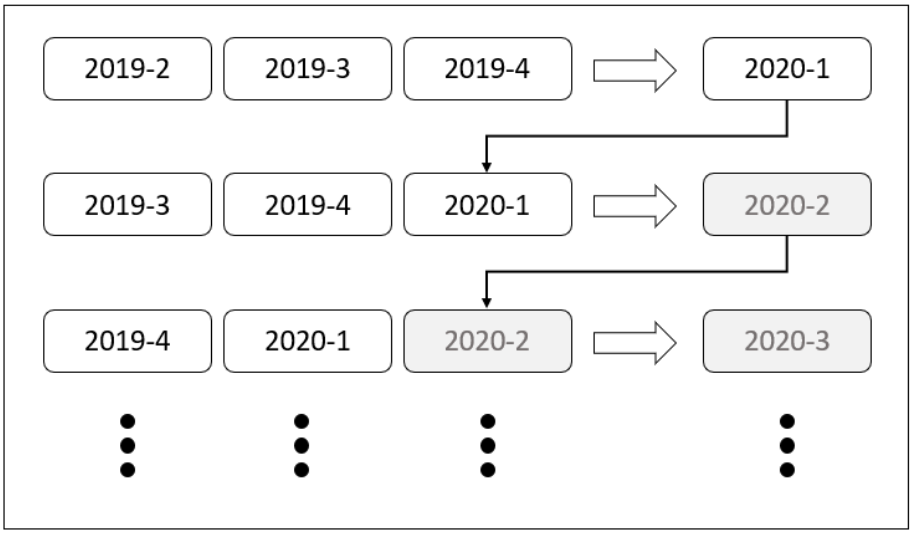
Predicción

Luego de ajustar los parámetros del modelo, se entrena y predice con todos los datos disponibles, obteniéndose un MAPE de 0.9053%. En la siguiente figura se puede observar el ajuste:



Con este modelo se pronosticaron los siguientes ocho trimestres. Para esto, se realizó una iteración de predicción con los últimos tres datos, es decir, se toman los

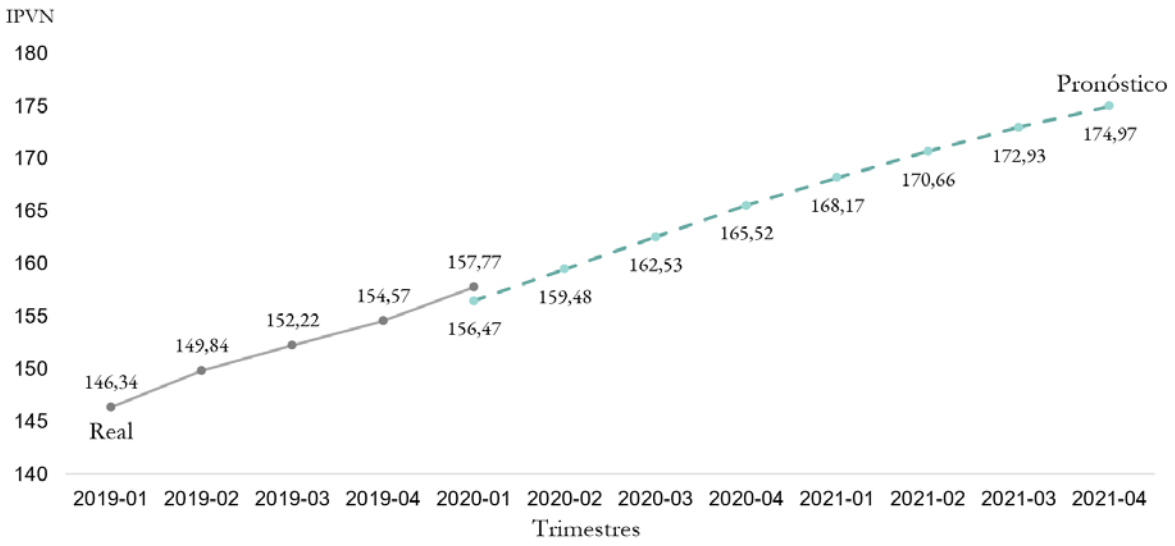
últimos registros disponibles para predecir el siguiente, que a su vez se vuelve en el último registro disponible, para la predicción que continúa. A continuación, se muestra de manera gráfica este proceso.



De esta manera, se obtuvieron los siguientes pronósticos:

Periodo	d_IPVN_MDE_Num	Proyeccion
2019-04	154.57	NaN
2020-01	157.77	156.466158
2020-02	NaN	159.478990
2020-03	NaN	162.529803
2020-04	NaN	165.523745
2021-01	NaN	168.166090
2021-02	NaN	170.657340
2021-03	NaN	172.927319
2021-04	NaN	174.972881

Se presenta la siguiente gráfica donde se muestra la predicción del IPVN desde 2020-2 hasta 2021-4, observándose que la predicción continúa la tendencia con la que venía el índice.



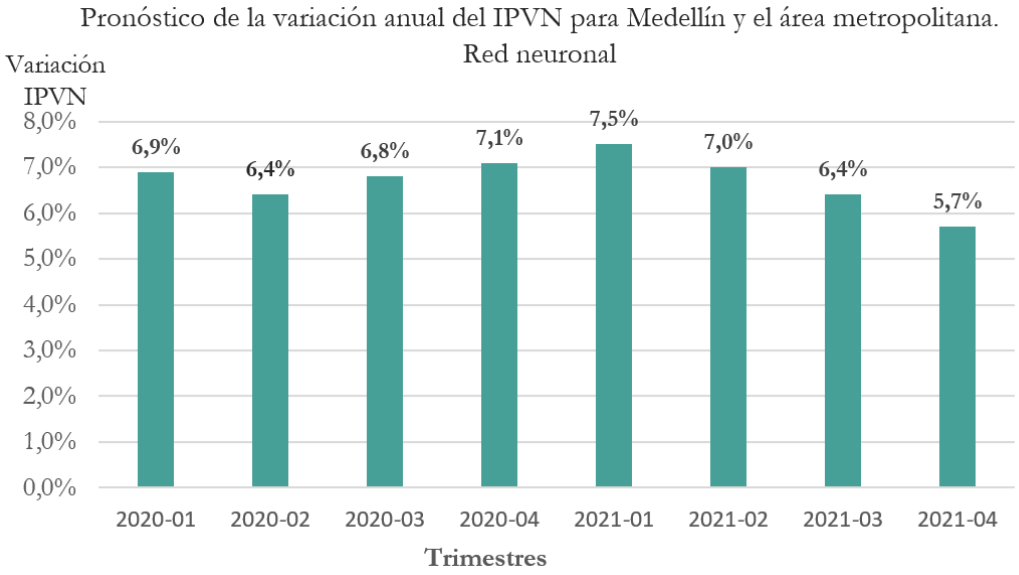
Evaluación de modelos

Con el fin de determinar el modelo de mejor pronóstico, se utilizó como criterio el error porcentual medio absoluto. Estos fueron los resultados obtenidos:

Modelo	MAPE
Regresión lineal multivariada	2.629
Red neuronal multivariada	2.28
Serie de tiempo univariante	0.617
Red neuronal univariante	0.531

En términos generales, las técnicas univariantes presentaron una mayor precisión. Tanto en la modelación multivariada como univariante, las redes neuronales presentaron mejores resultados frente a las técnicas estadísticas tradicionales, sin embargo, requieren de un mayor esfuerzo computacional y de una mayor experimentación, dado que no hay unas reglas claras para su determinación.

En definitiva, el modelo con mayor precisión, medido a través del MAPE, fue la red neuronal univariante. A continuación, se presentan los pronósticos obtenidos, a nivel de variación interanual.



De acuerdo con los resultados obtenidos con la red neuronal univariante, se espera que, en el cuarto trimestre de 2020, en promedio, el precio de la vivienda nueva en Medellín y su área metropolitana aumente un 7,1% con respecto al cuarto trimestre de 2019, mientras que para el 2021-4 se estima que esta variación interanual alcance el 5,7%. Sin embargo, para el año 2021, la tasa de crecimiento presentará una caída con respecto al 2020.

Recomendaciones y conclusiones

- Para el caso particular de la serie de Índice de Precios de la Vivienda Nueva - IPVN para Medellín y el Área metropolitana, el enfoque de modelación univariante ofrece una mayor precisión de pronóstico frente al enfoque multivariado. Sin embargo, no permite entender la incidencia de otras variables sobre su comportamiento, limitando el entendimiento del indicador a nivel de variables de oferta y demanda.

- En el caso particular de las redes neuronales, si bien presentan en ambos enfoques una mayor precisión frente a los métodos estadísticos tradicionales, requieren computacionalmente de más tiempo y rendimiento. Por otro lado, para la definición de parámetros, no hay claridad sobre unas reglas prácticas para llegar a estos sin necesidad de la experimentación.

- Si bien, se ofrecen resultados a nivel del índice, se exploraron modelos con las variaciones interanuales, obteniéndose precisiones bajas, y con un mayor esfuerzo computacional en el caso de las redes neuronales. Destacando que con el pronóstico del índice también se pueden obtener las variaciones, y con un nivel de precisión mayor.

- Una manera para que las redes neuronales a nivel multivariado obtengan una mayor precisión y menores costos computacionales, es trabajar con las variables obtenidas en la ingeniería de características, ya que el utilizar una gran cantidad de variables de entrada no contribuye a tener mejores resultados.

- Es importante validar en una regresión lineal no solo los supuestos que son necesarios para extender los resultados a inferencias, sino también la sensibilidad del modelo, para identificar si la cantidad de datos tiene un impacto sobre las variables escogidas; esto puede dar indicios de qué variables finalmente tomar para realizar la construcción del modelo, o cuestionar si esta metodología es la apropiada para el problema en estudio.

- Realizar un análisis de los resultados obtenidos en lo betas de un modelo de regresión lineal, ayuda a tener una mejor comprensión del impacto de las variables del modelo sobre la variable respuesta, este ejercicio invita a realizar una mejor comprensión de los resultados y tener un cuestionamiento de estos, retando al investigador a indagar por qué dichos resultados tienen sentido o no.

- Con los modelos desarrollados se identificaron algunas de las variables económicas que tienen un mayor efecto sobre el IPVN, es de resaltar principalmente el Índice de precio al consumidor de Medellín y el área metropolitana por parte de la regresión lineal y PIB valor a precio corriente base 2015 para el caso de las redes neuronales

Referencias bibliográficas

ATRIA Innovation. (22 de Octubre de 2019). ATRIA Innovation. Obtenido de Qué son las redes neuronales y sus funciones: https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/

Bressan, R. (28 de Febrero de 2020). Towards Data Science. Obtenido de Keras 101: A simple (and interpretable) Neural Network model for House Pricing regression: https://n9.cl/1du0

CAMACOL Cámara Colombiana de la Construcción. (2020). Obtenido de https://camacol.co/documentos/construccion-en-cifras

Castaño, E. (2020). Memorias del curso de Series de Tiempo. Universidad Eafit, Maestría de Finanzas.

DANE Departamento Nacional de Estadística. (2020). Cuentas Nacionales. Obtenido de https://bit.ly/33uYxcK

Jordi Torres.AI. (22 de Septiembre de 2019). Jordi Torres.AI. Obtenido de Redes Neuronales Recurrentes: https://torres.ai/redes-neuronales-recurrentes/

Kim, J. (7 de Agosto de 2019). Towards Data Science. Obtenido de Variable selection using LASSO: https://towardsdatascience.com/variable-selection-using-lasso-493ac2e5660d

Unipython. (2020). Unipython. Obtenido de Predicción con series temporales con LSTM, Redes neuronales recurrentes: https://unipython.com/prediccion-con-series-temporales-con-lstm-redes-neuronales-recurrentes/

Wei, W. (2006). Time Series Analysis: Univariate and Multivariate Methods. Pearson Education.

Repositorio

La información de este trabajo se encuentra disponible en el repositorio de GitHub https://github.com/SusanaLondono/PI2_CDDDS