Modelación Estadística II
¿Cuál casa comprar?

Iván Vega Gutiérrez

Centro de Investigación en Matemáticas A.C.
Unidad Aguascalientes
E-mail: ivan.vega@cimat.mx

I. Introducción

En el sector inmobiliario, un problema de especial interés es el de determinar el precio mas adecuado de compra o de venta de casas-habitación. Los dueños de los inmuebles (o las compañías inmobiliarias) suelen tomar en cuenta diferentes aspectos y condiciones para definir dicho precio. Una forma de estandarizar este mercado, sería el medir el efecto que tienen algunas de las características del inmueble sobre el precio que se estipula. Se llevó a cabo un estudio piloto en este sentido con el fin de obtener información que permita, posteriormente, diseñar un estudio a gran escala para responder al objetivo de estandarización. En el estudio piloto se seleccionaron 26 casas en venta y se registró las siguientes variables:

- RECAM: Número de recámaras.
- AREA: Metros cuadrados de construcción.
- CHIM: Variable dicotómica que indica si la casa tiene o no chimenea.
- CUARTOS: Número de cuartos que tiene la casa.
- CONTRAV: Variable dicotómica que indica si la casa tiene o no contra ventanas.
- LONGFR: Longitud del frente en metros.
- BAÑOS: El número de baños que tiene la casa.
- CON: Variable dicotómica que indica si la construcción es de ladrillo o tiene madera.
- COCHERA: Número de autos.
- COND: Variable dicotómica que indica si la casa necesita arreglos mayores.
- ZONA: Variable categórica que indica la zona de la ciudad en donde se localiza la casa.
- PRECIO: En miles de pesos.

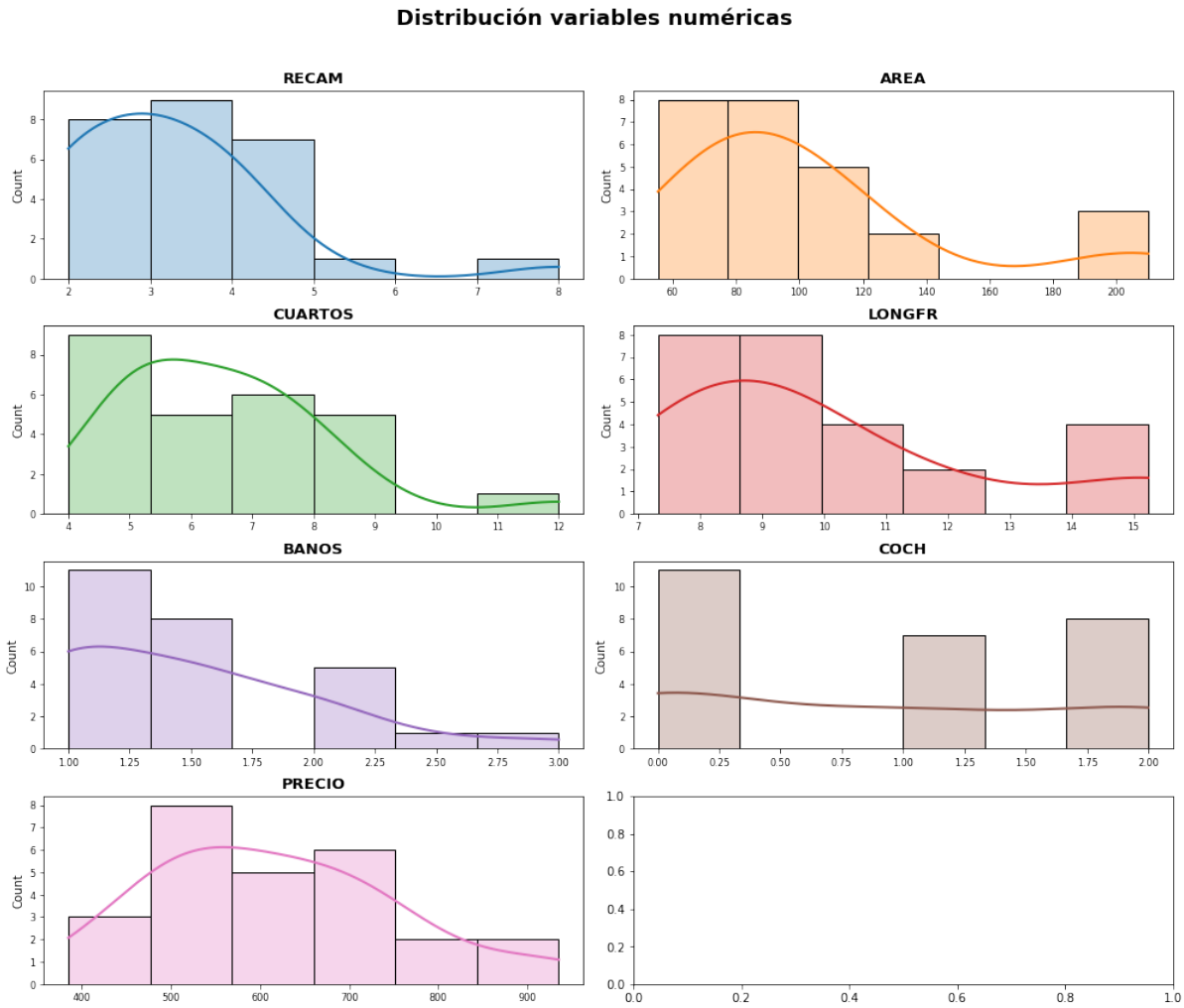

Figura 1: Histogramas para las variables cuantitativas

II. Análisis exploratorio de datos

Para el análisis exploratorio de datos se utilizó la versión 3.8.8 de Python, en específico se usó la librería pandas profiling, la cual es una herramienta muy útil que nos permite tener una visualización general de los datos, la forma en la que se distribuyen, las posibles relaciones que existen entre las variables, los principales estadísticos, etc.

La base de datos está conformada por 26 observaciones y 12 variables. Nuestra variable dependiente es la variable PRECIO. Dentro de las variables se encuentran tanto de tipo categórico como cuantitativo. En la figura (1) se muestran los histogramas de las variables numéricas. Por otro lado en (2) se muestran la relación de las variables categóricas con respecto a la variable PRECIO.

A partir de las figura (1) podemos ver que la variable PRECIO pareciera tener una distribución normal y que las otras variables cuantitativas tienen un comportamiento similar, el comportamineto es asimétrico, se concentra en mayor cantidad en los valores mínimos de la variable y conforme va aumentando la probabilidad va disminuyendo, es decir, presentan una asimetría positiva, lo cual se traduce que hay una mayor cantidad inmuebles con pocos cuartos, baños, recamaras, chocheras, etc y mientras mayor es el número de cuartos, longitud frontal, baños, área, etc, menos los inmuebles que

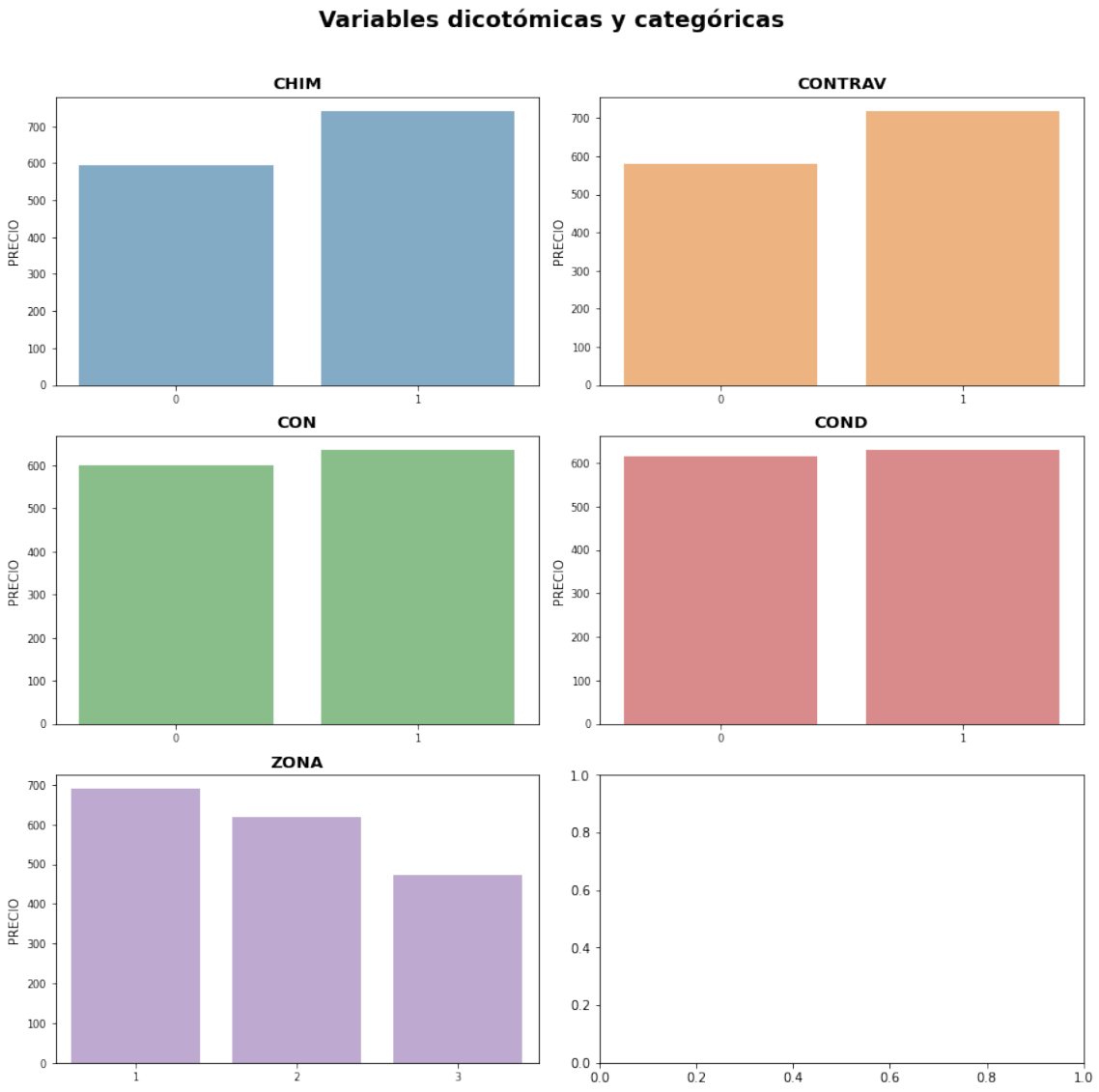

Figura 2: Relación entre las variables categóricas y el PRECIO

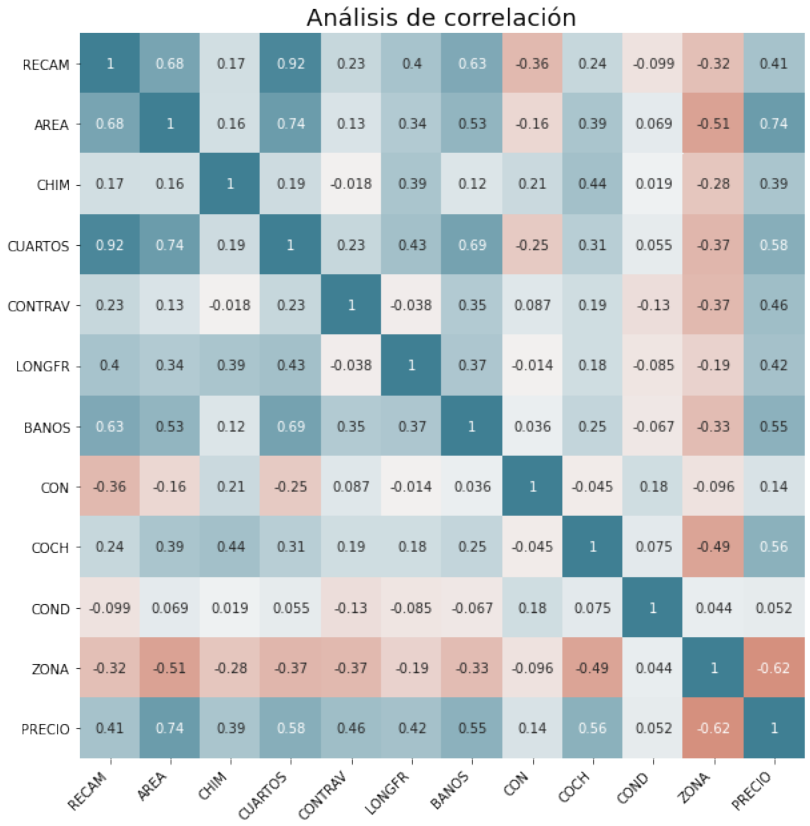

Figura 3: Matriz de correlación

poseen estas características.

Por otro lado, en la figura (2) se observa que las características de las casas influyen directamente en el precio, por ejemplo, las casas que tienen chimenea o contraventana parecen ser más costosas. Además, las casas de ladrillo parecieran ser ligeramente más costosas que las de madera, de igual manera, las casas que necesitan arreglos mayores son un poco más económicas, sin embargo no pareciera ser un factor que influya mucho en el precio. Y por último, con respecto a las zonas, se puede afirmar que la zona más cara es la 1, seguida de la zona 2 y por último la zona 3, que pareciera ser la zona más económica.

Asimismo, el valor promedio de una casa es de 617 mil pesos aproximadamente, y la desviación estándar de la variable PRECIO es 138.3970431. También se puede observar que la casa más económica oscila entre los 385 mil pesos, mientras que la casa más costosa ronda entre los 935 mil pesos.

III. Análisis de correlación

Como se vio en la sección previa, pareciera que hay algunas variables que no influyen significativamente en el precio como lo son la condición y el tipo de material, mientras que hay otras que parecieran tener un gran impacto en el precio como el área. Para poder tener una mejor interpretación, en (3) se muestra el gráfico de correlación.

De la matriz de correlación, se observa que las variables que tienen una correlación positiva con el PRECIO son: ÁREA, CUARTOS, BAÑOS y COCHERA. Mientras que la variable ZONA guarda una correlación negativa con el precio bastante interesante. Otro dato interesante, aunque bastante obvio es que la variable CUARTOS y RECAM tienen una correlación fuerte positiva, lo cual es importante

a tomar en cuenta para que no existan errores de colinealidad posteriormente.

IV. Análisis de Varianza

Para poder afirmar los hallazgos encontrados en el análisis descriptivo de los datos y de la matriz de correlación, procedemos a hacer una prueba ANOVA, para ver si las variables categóricas influyen de manera significativa con respecto al precio.

	sum_sq	df	F	PR(>F)
CHIM	32561.890857	1.0	2.851967	0.106794
CONTRAV	40406.138048	1.0	3.539014	0.074580
CON	53.298125	1.0	0.004668	0.946206
COND	5294.632579	1.0	0.463736	0.503692
ZONA	68698.164788	1.0	6.017000	0.023470
Residual	228346.898789	20.0	NaN	NaN

De los resultados obtenidos de la prueba ANOVA, se puede observar que las únicas variables significativas con respecto al precio son CONTRAV y ZONA. Observemos también que la variable CHIMENEA es la que pareciera ser significativa después de contraventana y la zona, sin embargo, el valor F = 2.85 no supera el valor crítico para una significancia $\alpha = 0.05$, por lo tanto, la variable CHIM no es una variable significativa que influya sobre PRECIO.

V. Análisis de Regresión Lineal

Ahora bien, hallemos un modelo de regresión lineal múltiple que nos permita predecir el valor del PRECIO con respecto de las demás variables, a pesar de que el análisis de varianza nos sugiere tomar en cuenta solo las variables ZONA y CONTRAV, como un primer modelo, tomaremos en cuenta todas las variables y de esa forma veremos que tan significativas son las variables con respecto a la variable PRECIO, para poder ajustar mejor el modelo con una nueva elección de variables independientes y posteriormente obtener nuevos resultados.

OLS Regression Results

Dep. Variable:	PRECIO	R-squared:	0.938
Model:	OLS	Adj. R-squared:	0.882
Method:	Least Squares	F-statistic:	16.50
Date:	Thu, 12 May 2022	Prob (F-statistic):	6.30e-06
Time:	01:19:38	Log-Likelihood:	-128.34
No. Observations:	26	AIC:	282.7
Df Residuals:	13	BIC:	299.0
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	189.1762	63.655	2.972	0.011	51.659	326.694
C(CHIM)[T.1]	39.9523	35.788	1.116	0.284	-37.364	117.269
C(CONTRAV)[T.1]	103.714	25.292	4.101	0.001	49.075	158.354
C(CON)[T.1]	30.4332	25.905	1.175	0.261	-25.530	86.397
C(COND)[T.1]	2.3771	26.228	0.091	0.929	-54.284	59.039

C(ZONA)[T.2]	60.8894	31.309	1.945	0.074	-6.750	128.529
C(ZONA)[T.3]	-15.0558	31.928	-0.472	0.645	-84.031	53.920
RECAM	-56.3979	24.084	-2.342	0.036	-108.429	-4.367
AREA	2.0490	0.374	5.478	0.000	1.241	2.857
CUARTOS	23.4741	19.640	1.195	0.253	-18.955	65.903
LONGFR	10.8844	4.567	2.383	0.033	1.019	20.750
BANOS	26.0430	27.480	0.948	0.361	-33.325	85.411
COCH	40.3816	17.091	2.363	0.034	3.460	77.304

Omnibus:	0.546	Durbin-Watson:	1.544
Prob(Omnibus):	0.761	Jarque-Bera (JB):	0.635
Skew:	-0.281	Prob(JB):	0.728
Kurtosis:	2.479	Cond. No.	830.

Observemos que el coeficiente de determinación de nuestro modelo es $R^2 = 9.38$, por lo que el modelo es capaz de explicar el 93.8 % de la variabilidad del precio, sin embargo, este no es un buen indicador ya que el número de variables independientes es bastante grande, sin embargo, podemos observar que el coeficientes de determinación ajustado es de 0.882, el cual afortunadamente sigue siendo alto. Además, el p-valor del modelo es significativo $6.30e-06$, por lo tanto, podemos concluir que el modelo es significativo y que al menos una de las variables independientes está fuertemente relacionada con el valor del precio.

En general, se observan los p-valores para cada variable independiente. De aquí, se puede ver que la mayoría de las variables no son significativas. En particular la variable CHIM no es una variable significativa para el modelo, el coeficiente asociado que aparece de la variable chimenea se puede interpretar como, la forma en la que afecta el precio de una casa si esta tiene o no chimenea es de 39.9523 mil pesos. Las variables significativas para el modelo con un nivel de significacia $\alpha = 0.05$ únicamente son : CONTRAV, RECAM, AREA, LONGFR Y COCH.

VI. Análisis de residuos

En (4) se muestran las respectivas gráficas para el análisis de residuos. La gráfica superior izquierda muestra la distribución de los residuos en cada observación, mientras que la gráfica superior derecha muestra los residuos con respecto a la variable predictora, de estas gráficas se puede observar que los residuos se reparten una forma más o menos simétrica .

La gráfica inferior izquierda es una gráfica Q-Q plot, en la cual podemos observar que los residuos presentan un comportamiento lineal, lo cual nos sugiere que posiblemente tengan una distribución normal.

Sin embargo, la última gráfica muestra el histograma de los residuos, el cual a simple vista dista de tener una distribución normal. Para asegurar estas observaciones, procedemos a realizar una prueba de Shapiro-Wilks, la cual nos da un p valor de 0.48 aproximadamente, por lo tanto no se rechaza la hipótesis de que los residuos siguen una distribución normal. Asimismo, se realizó una prueba de White, la cual arrojó que los residuos presentan homocedasticidad.

VII. Métodos Forward, Backward y Sepwise

Los métodos Forward sirven para seleccionar variables significativas en el modelo de regresión lineal. El método Forward inicia sin ninguna variable signficativa y en cada paso se van agregando variables significativas, por otro lado, el método Backward inicia con todas las variables tomándolas como significativas y en cada paso va eliminando variables que no son significativas, mientras que el

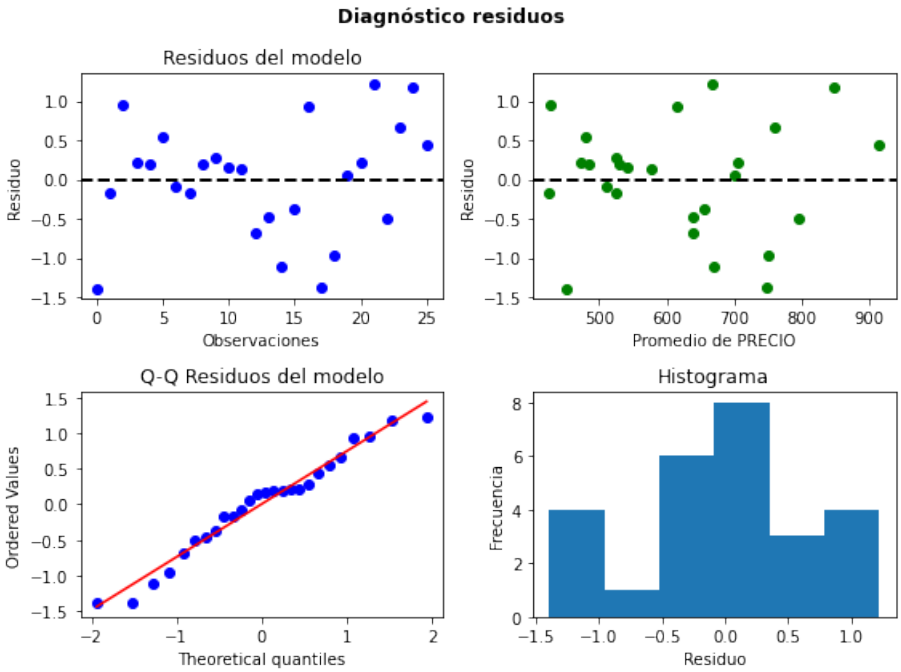

Figura 4: Gráficas de residuos

método Stepwise es una combinación de ambas, al igual que el método Forward inicia sin variables y en cada paso va agregando variables significativas, sin embargo, en cada paso verifica que las variables que se agreguen realmente sean signficativas, es decir combina el método Backward después de agregar alguna variable.

Para implementar estos tres métodos se utilizó la versión 4.1.0 de R, los resultados de los métodos coincidieron, dando como resultado la selección de las variables:

- RECAM
- AREA
- CHIM
- CUARTOS
- CONTRAV
- LONGFR
- COCH

Al seleccionar estas variables como significativas el coeficiente de determinación ajustado fue de 0.8593 con un p-valor de $1.006e-07$. El cual da un mejor resultado que el modelo inicial que contempla todas las variables como significativas, sin embargo, en el análisis anterior concluimos que la variable CHIM no es significativa, por lo que se propone un nuevo modelo en el cual no se tome en cuenta las variables CHIM y CUARTOS, para comprobar la efectividad de estos métodos.

VIII. Modelos de Machine Learning

En esta sección se imlementaron tres modelos de Machine Learning:

1. K-Nearest-Neighbor
2. Random Forest
3. Regresión Lineal (Ridge y Laso)

El porcentaje de datos para el entrenamiento fueron del 80 % y el 20 % para los datos de prueba. En la figura (5) se muestra la evolución del error para los métodos K-Nearest-Neighbor y Regresión lineal respectivamente.

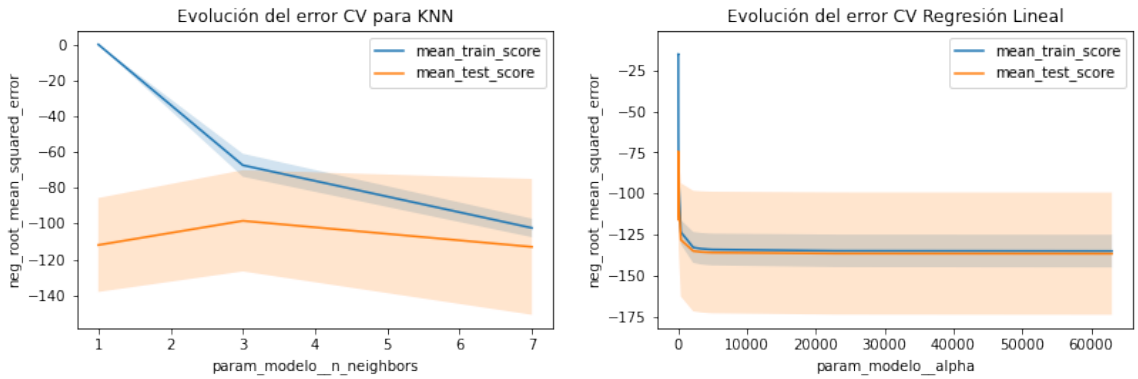

Figura 5: Evolución de los errores

En la tabla (1) se muestran los errores de cada método implementado.

Método	Error
KNN	74.71067651403771
Random Foest	80.26008368630909
Regresión Lineal	77.45566485608906

Tabla 1: Errores de los métodos de ML

De la talba de errores, podemos observar que el método con el menor error es el de KNN, seguido de Regresión Lineal y por último Random Forest. Cabe señalar que para implementar cualquier método de machine learning, es necesario un gran número de obsrvaciones, sin embargo, nuestra base de datos solamente cuenta con 26 observaciones, por lo tanto, no es adecuado ningún método de machine learning, a menos que se obtengan más observaciones, esto con el fin de poder hacer predcciones más precisas.

IX. Conclusiones

De todos los métodos propuestos, solo queda comparar el modelo que se obtuvo aplicando los métodos de Forward, Backward y Stepwise(Modelo 2) y el modelo propuesto con las variables significativas (Modelo 3) que se obtuvo con el análisis estadístico. En la tabla (2) se muestran los tres modelos de regresión lineal propuestos.

Observemos que el modelo que tiene un mejor rendimiento es el modelo propuesto de los métodos Forward, Backward y Stepwise. En la figura (6) se observa que el modelo 2 presenta una mejor normalidad en los residuos, lo cual es importante para poder aplicar de manera adecuada el modelo de regresión lineal. Por lo tanto, se puede concluir que tanto los modelos 2 y 3 son buenos a comparación de tomar todas las variables como significativas(Modelo 1).

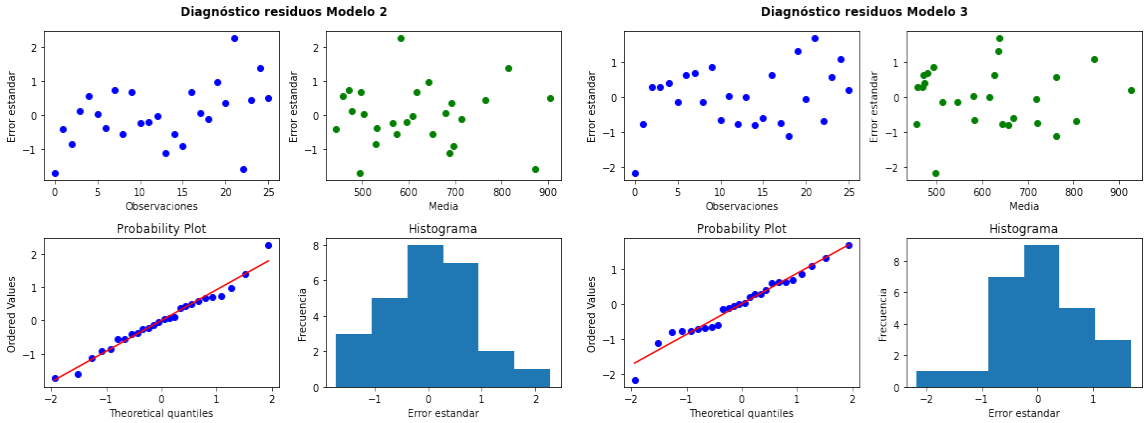

Figura 6: Comparación de residuos del modelo 2 y 3

X. Bibliografía

https://www.cienciadedatos.net/

Modelo	coeficiente de determinación ajustado	p-valor
Modelo 1	0.882	6.30e-06
Modelo 2	0.791	3.72e-07
Modelo 3	0.859	1.01e-07

Tabla 2: Comparación de modelos