Predicción de Precios de Casas

Universidad de Costa Rica
Escuela de Matemática

Departamento de Matemática y Ciencias Actuariales
CA0305: Herramientas de Ciencia de Datos II

Integrantes:
Ashley Arrieta Padilla - C00753
Sofía Bocker Brenes - C11102
Bryan Campos Vega - C01654
Naydelin Hernández Vargas - C03795
Dixon Montero Hernández - B99109

Profesor:
Prof. Luis Alberto Juárez Potoy

10 de julio de 2024

1 Introduccón

El valor de las viviendas es una métrica crucial en la economía, afectando tanto a individuos como a la salud económica de una región. Una predicción precisa facilita la toma de decisiones, la transparencia del mercado y la identificación de tendencias y oportunidades de inversión.

El objetivo de este proyecto es desarrollar una metodología para predecir el valor de las viviendas en diversas zonas utilizando métodos de regresión en Python. Se aplican técnicas de análisis de datos y modelado predictivo para identificar la relación entre el valor de las viviendas y otras variables. Desarrollar modelos predictivos permite entender cómo estas variables influyen en los precios y predecir el valor de nuevas propiedades. Esto es esencial para los actores del mercado inmobiliario que buscan maximizar su retorno de inversión y minimizar riesgos.

La importancia de este trabajo radica en mejorar la precisión y confiabilidad de las predicciones de precios de viviendas. Además, la metodología desarrollada puede servir como base para futuras investigaciones en análisis de datos inmobiliarios.

Al aplicar diferentes modelos de regresión y comparar su desempeño, se busca identificar el modelo que mejor se ajusta a los datos disponibles. Este enfoque no solo optimiza la precisión de las predicciones, sino que también proporciona una comprensión más profunda de las dinámicas del mercado inmobiliario, lo cual es fundamental para la toma de decisiones informadas en el sector.

2 Metodología

Para este proyecto se ha desarrollado una metodología para predecir el valor de viviendas en diversas zonas utilizando métodos de regresión en Python. Nuestro enfoque se basa en el uso de técnicas de análisis de datos y modelado predictivo para establecer una relación entre el valor de las viviendas y una serie de variables explicativas seleccionadas.

Para llevar a cabo este análisis, se han seleccionado diversas variables que se consideran influyentes en el valor de las propiedades. Estas variables pueden incluir, pero no se limitan a, la ubicación geográfica, el tamaño de la propiedad, el número de habitaciones y baños y la antigüedad de la vivienda. La metodología aplicada permite no solo entender cómo cada una de estas variables impacta el valor de las viviendas, sino también predecir el precio de nuevas propiedades con base en sus características.

Se utilizaron 4 modelos para la regresión de los datos para la predicción del valor de las viviendas:

2.1 Regresión lineal simple

La regresión lineal es un modelo matemático que describe la relación entre varias variables. Los modelos de regresión lineal son un procedimiento estadístico que ayudan a predecir el futuro. En este caso, se utiliza para predecir los precios de las viviendas en las zonas seleccionadas de la base de datos.

Debido a su capacidad para transformar datos, pueden utilizarse para simular una amplia gama de relaciones, y gracias a su forma, que es más simple que la de las redes neuronales, sus parámetros estadísticos se analizan y comparan con facilidad, lo que permite que se les extraiga información valiosa. (Saavedra, 2023)

2.2 Modelo polinomial (Ridge)

La regresión Ridge modifica la función de coste para minimizar la complejidad del modelo. Se mide como la suma de los cuadrados de los coeficientes. Es decir, como ejemplo de regresión lineal,

una función del modelo sería $y = mx + n$, con Ridge la función de coste sería $\alpha * (n^2 + m^2)$. El hiperparámetro α (alpha) controla cuánta regularización queremos aplicar a nuestro modelo. Si es 0, entonces regresión Ridge se convierte en regresión lineal. Cuando α se hace más grande todos los parámetros acaban con valores cercanos a 0 (DELTA, 2020).

2.3 Modelo polinomial

La regresión polinomial es una forma sencilla de adaptar un modelo de regresión lineal a estructuras más complejas. Para ello solo hace falta crear nuevos parámetros elevando al cuadrado, cubo o cualquier otra cifra las variables ya existentes (DELTA, 2020).

La ecuación de un modelo de regresión polinomial de grado n se puede expresar como:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n + \epsilon$$

2.4 Regresión elástica Net

Elastic Net es un modelo de regresión lineal que normaliza el vector de coeficientes con las normas L1 y L2. Esto permite generar un modelo en el que solo algunos de los coeficientes sean no nulos, manteniendo las propiedades de regularización de Ridge. La función de coste es equivalente a:

$$RSS_{elastic\,net} = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \alpha \left(\lambda \sum_{j=1}^{p} \beta_j^2 + (1 - \lambda) \sum_{j=1}^{p} |\beta_j| \right)$$

Este método utiliza la clase Python sklearn.linear_model.ElasticNet para estimar los modelos de regresión lineal regularizada para una variable dependiente en una o más variables independientes. La regularización combina las penalizaciones L1 (Lasso) y L2 (Ridge). Cuando se ajusta un único modelo o se utiliza la validación cruzada para seleccionar la relación de penalización y/o alfa, se puede utilizar una partición de datos reservados para estimar el rendimiento fuera de la muestra. (IBM, 2021)

Además de ajustar un modelo con valores especificados de la razón de la penalización L1 y el parámetro de regularización alfa, la red elástica lineal puede mostrar un trazo de valores de coeficiente para un rango de valores alfa para una razón determinada, o facilitar la elección del valor de hiperparámetros a través de la validación cruzada k-fold en cuadrículas de valores especificadas. (IBM, 2021)

2.5 Regresión modelo Lasso

El modelo de Lasso es un modelo lineal que penaliza el vector de coeficientes añadiendo su norma L1 (basada en la distancia Manhattan) a la función de coste:

$$RSS_{lasso} = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Lasso tiende a generar ´coeficientes dispersos': vectores de coeficientes en los que la mayoría de ellos toman el valor cero. Esto quiere decir que el modelo va a ignorar algunas de las características predictivas, lo que puede ser considerado un tipo de selección automática de características. Al incluir menos variables suponemos un modelo más sencillo de interpretar (Burrueco, 2022).

2.6 Idea del código

Clase: `Data`

Esta clase se encarga de gestionar y preparar los datos para el modelado. Tiene los siguientes métodos:

- **`__init__`**: Inicializa la clase con los atributos `data`, `descrip_entrenamiento`, `descrip_prueba`, `obj_entrenamiento` y `obj_prueba`.
- **`cargar_data`**: Carga datos desde un archivo CSV a un `DataFrame` de pandas.
- **`get_data`**: Retorna las primeras filas del `DataFrame`.
- **`__str__`**: Retorna una cadena de texto con las dimensiones del `DataFrame` y la cantidad de valores nulos por columna.
- **`eliminar_columnas`**: Elimina las columnas especificadas del `DataFrame`.
- **`convertir_columnas_descriptivas_a_numericas`**: Convierte las columnas descriptivas especificadas a tipo numérico.
- **`eliminar_unknown`**: Elimina las filas donde la columna especificada tenga el valor `'Unknown'`.
- **`imputar_por_grupo`**: Imputa valores faltantes en una columna basada en las estadísticas (por ejemplo, la media) de los grupos definidos por otras columnas.
- **`preparar_data`**: Prepara los datos para el entrenamiento, dividiéndolos en conjuntos de entrenamiento y prueba, y escalando las variables numéricas.

Clase `Modelo` (hereda de `Data`)

Esta clase se encarga de implementar y evaluar varios modelos de regresión. Tiene los siguientes métodos:

- **`Modelo_Ridge`**: Implementa una regresión Ridge con búsqueda en cuadrícula para seleccionar el mejor parámetro de regularización α. Imprime las métricas de evaluación (RMSE, R^2, R^2 ajustado y MAE) para el conjunto de prueba.
- **`Modelo_regresion_polinomial`**: Implementa una regresión polinomial. Imprime las métricas de evaluación (RMSE, R^2, R^2 ajustado y MAE) para el conjunto de prueba.
- **`Modelo_regresion_elastic_net`**: Implementa una regresión Elastic Net con validación cruzada repetida. Imprime las métricas de evaluación (RMSE, R^2, R^2 ajustado y MAE) para el conjunto de prueba.
- **`Modelo_regresion_lineal`**: Implementa una regresión lineal simple. Imprime las métricas de evaluación (RMSE, R^2, R^2 ajustado y MAE) para el conjunto de prueba.
- **`Modelo_Lasso`**: Implementa una regresión Lasso con búsqueda en cuadrícula para seleccionar el mejor parámetro de regularización α. Imprime las métricas de evaluación (RMSE, R^2, R^2 ajustado y MAE) para el conjunto de prueba.

Clase `Graficos` (hereda de `Data`)

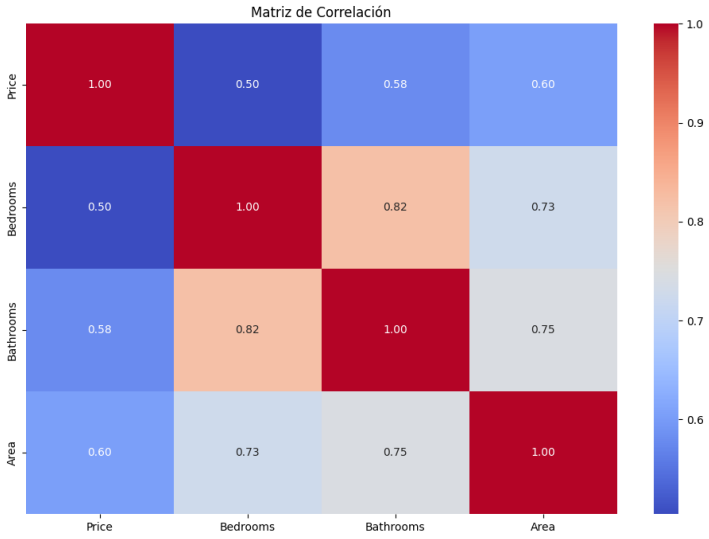
Esta clase se encarga de generar gráficos para visualizar los datos. Tiene los siguientes métodos:

- **`__init__`**: Inicializa la clase con los atributos `data`, `descrip_entrenamiento`, `descrip_prueba`, `obj_entrenamiento` y `obj_prueba`.
- **matriz_correlacion**: Grafica una matriz de correlación entre una variable objetivo y las demás variables cuantitativas.
- **grafico_dispersion**: Muestra un gráfico de dispersión comparando dos variables cuantitativas.
- **grafico_boxplot**: Muestra un gráfico boxplot para una columna específica de la base de datos.
- **histograma**: Muestra un histograma de una columna específica con una cantidad determinada de *bins*.

En los métodos de la regresión polinomial ridge, elastic net y lasso se hizo uso de GridSearchCV con el propósito de poder encontrar el *alpha* que más se ajuste al modelo y utilizar para su aplicación. Además, a cada uno de los tipos de regresiones se le calcularon diferentes métricas de error, con esa finalidad se utilizaron los métodos importados mean_squared_error, r2_score, mean_absolute_error. Luego, se llamó el módulo con todas las clases y métodos en un archivo main.ipynb para poder correrlos y obtener los resultados, los cuales se discutirán en la siguiente sección:

3 Resultados

Figura 1: Matriz de Correlación con la Variable Objetivo



Fuente: Elaboración propia

En la figura 1 se muestra la matriz de correlación entre la variable objetivo y las demás variables. En esta, se puede observar que *Area* presenta el valor más alto con respecto a *Price* mientras que *Bedrooms* y *Bathrooms* poseen el mismo valor, que a la vez es el más bajo.

Las métricas de desempeño de modelos son herramientas fundamentales en el aprendizaje automático para evaluar y cuantificar la efectividad de los modelos de predicción y clasificación. Estas métricas nos permiten determinar qué tan bien un modelo realiza predicciones o clasifica datos.(González, 2008) Se han llevado a cabo predicciones con los datos seleccionados y se han calculado diversas métricas de evaluación para medir el desempeño de los modelos. A continuación, se proporciona una explicación general de cada métrica a analizar

- RMSE (Root Mean Squared Error): Es conocida también como la desviación cuadrática media. Mide la magnitud promedio de los errores y se ocupa de las desviaciones del valor real.(Vitalflux, s.f.) Cuanto menor sea el RMSE, mejor será el ajuste del modelo a los datos.
- R-squared (Coeficiente de Determinación): Esta métrica indica la proporción de la variabilidad en la variable dependiente que puede ser explicada por las variables independientes del modelo. Un R- squared más alto indica un mejor ajuste del modelo. (Vitalflux, s.f.)
- R-squared Ajustado: Esta métrica ajusta el R-squared estándar teniendo en cuenta el número de predictores en el modelo y el tamaño de la muestra. (Vitalflux, s.f.)
- MAE (Mean Absolute Error): Esta métrica mide el error promedio absoluto entre las predicciones y los valores reales. Al igual que con el RMSE, un MAE más bajo indica un mejor desempeño del modelo.(Vitalflux, s.f.)

A continuación se analizaran los resultados obtenidos para cada modelo, basandonos en la explicación de las métricas encontradas en: (RStudio, s.f.)

3.1 Modelo de regresión lineal

Las métricas de evaluación obtenidas por el Modelo de regresión lineal en los datos son las siguientes:

Métrica	**Valor**
RMSE (Prueba)	4812445.43
R-squared (Prueba)	0.4528
R-squared ajustado (Prueba)	0.4518
Error Absoluto Medio (MAE) (Prueba)	2601778.68

Cuadro 1: Resultados de evaluación del modelo Regresión lineal

- **RMSE (Prueba):** Permite evaluar qué tan bien generaliza el modelo a nuevos datos. El valor del RMSE en el conjunto de prueba es bastante alto, se obtuvo un valor de 5301587.60. Esto nos indica que, en promedio, las predicciones del modelo de regresión lineal están bastante alejadas de los valores reales en el conjunto de datos de entrenamiento.
- **R-squared (Prueba):** Un valor de 0.4313 indica que el modelo explica el 43.13 % de la variabilidad de los datos de entrenamiento. A comparación de los otros modelos analizados este valor es bastante bajo. Un valor cercano a 1 sería ideal, ya que así el modelo explicaría casi toda la variabilidad en los datos

- **R-squared ajustado:** Un R-squared ajustado de 0.4302 indica que elmodelo explica aproximadamente el 43.02 % de la variabilidad en la variable de respuesta, considerando la cantidad y relevancia de los predictores incluidos. Es común que el R-squared ajustado sea ligeramente más bajo que el R-squared estándar. (Cosio, 2021)
- **Error Absoluto Medio (MAE) (Prueba):** Un MAE de 2618741.59 sugiere que el modelo tiene un nivel medio de precisión en las predicciones en el conjunto de prueba

3.2 Modelo polinomial (Ridge)

Las métricas de evaluación obtenidas por el Modelo polinomial (Ridge) para los datos son las siguientes:

Métrica	**Valor**
RMSE (Prueba)	4705248.58
R-squared (Prueba)	0.4769
R-squared ajustado (Prueba)	0.4688
Error Absoluto Medio (MAE) (Prueba)	2432840.60

Cuadro 2: Resultados de evaluación del modelo Polinomial (Ridge)

- Dado el valor de **RMSE** de 5116744.50 , podemos decir que el modelo tiene una desviación promedio considerable en sus predicciones, es decir, el modelo Ridge con transformación polinomial de grado 2 tiene un margen de error significativo en sus predicciones.
- **R-squared :** Aproximadamente el 47.02 % de la variabilidad en la variable de respuesta (cuánto cambia la variable de interés en función de las variables predictoras) puede ser explicada por este modelo. Un valor cercano a 1 sería ideal, ya que así el modelo explicaría casi toda la variabilidad en los datos.
- **R-squared ajustado:** Un R-squared ajustado de 0.4620 indica que el modelo explica aproximadamente el 46.20 % de la variabilidad en la variable de respuesta, considerando la cantidad y relevancia de los predictores incluidos. Es común que el R-squared ajustado sea ligeramente más bajo que el R-squared estándar. (Cosio, 2021)
- **Error Absoluto Medio (MAE) :** Un valor de MAE de 2460836.13 indica que, en promedio, las predicciones del modelo se desvían de los valores reales en aproximadamente 2460836.13 unidades. Este valor nos da una idea de la magnitud de los errores que el modelo comete al hacer predicciones sobre los datos analizados.

3.3 Modelo de Regresión Polinomial.

Métrica	**Valor**
RMSE (Prueba)	4701489.73
R-squared (Prueba)	0.4778
R-squared ajustado(Prueba)	0.4696
Error Absoluto Medio (MAE) (Prueba)	2433018.77

Cuadro 3: Resultados de evaluación del modelo Regresión Polinomial

- **RMSE:** Podemos notar que el RMSE es alto para el conjunto de prueba, esto indica que las predicciones del modelo tienen una desviación significativa respecto a los valores reales.Un RMSE más bajo en el conjunto de prueba sería deseable, ya que indicaría que el modelo está haciendo predicciones más precisas en datos que no ha visto durante el entrenamiento (RStudio, s.f.)

- **R-squared :** El valor de 0.4702 sugiere que el 47.02 % de la variabilidad en los datos de entrenamiento es explicada por el modelo. Esto indica un ajuste moderado.Un valor cercano a 1 sería ideal, ya que así el modelo explicaría casi toda la variabilidad en los datos.(Cosio, 2021)

- **R-squared ajustado:** El R-squared ajustado de 0.4619 es ligeramente menor que el R-squared normal. Es decir, teniendo en cuenta el número de predictores, el modelo todavía explica el 46.19 % de la variabilidad en los datos de prueba. (Cosio, 2021)

- **Error Absoluto Medio (MAE):** Un MAE de 2462043.02 indica que, en promedio, las predicciones del modelo difieren de los valores reales en aproximadamente 2,392,013 unidades. El valor obtenido es bastante alto, lo que indica una desviación significativa en las predicciones.

3.4 Modelo Elastic net

Métrica	**Valor**
RMSE (Prueba)	4984013.07
R-squared (Prueba)	0.4131
R-squared ajustado (Prueba)	0.4120
Error Absoluto Medio (MAE) (Prueba)	2619460.74

Cuadro 4: Resultados de evaluación del modelo Elastic Net

- **RMSE (Prueba):** El valor del RMSE en el conjunto de prueba es 5134160.34.podemos decir que el modelo tiene una desviación promedio considerable en sus predicciones, es decir, el modelo Elastic Net tiene un margen de error significativo en sus predicciones.

- **R-squared (Prueba):** El coeficiente de determinación (R-squared) en el conjunto de prueba es 0.3893. Este valor indica que aproximadamente el 38.93 % de la variabilidad en los datos de prueba puede ser explicada por el modelo. Un valor de R-squared más cercano a 1 es deseable, ya que indica un mejor ajuste del modelo a los datos.(RStudio, s.f.)

- **R-squared ajustado (Prueba):** El R-squared ajustado en el conjunto de prueba es 0.3882. considerando la cantidad y relevancia de los predictores incluidos. Es común que el R-squared ajustado sea ligeramente más bajo que el R-squared estándar.

- **Error Absoluto Medio (MAE) (Prueba):** Un valor de MAE de 2605883.10 indica que, en promedio, las predicciones del modelo se desvían de los valores reales en aproximadamente 2605883.10 unidades. Este valor nos da una idea de la magnitud de los errores que el modelo comete al hacer predicciones sobre los datos analizados.

3.5 Modelo Lasso

Métrica	Valor
RMSE (Prueba)	4812423.72
R-squared (Prueba)	0.4528
R-squared ajustado (Prueba)	0.4518
Error Absoluto Medio (MAE) (Prueba)	2601676.36

Cuadro 5: Resultados de evaluación del modelo Lasso

- **RMSE:** Un RMSE de 5030230.97 señala que en promedio las predicciones del modelo están a una distancia de aproximadamente 5030230.97 unidades de los valores reales. Este es un valor bastante alto, lo que sugiere que las predicciones tienen una desviación significativa de los valores reales.
- **R-squared:** Un valor de 0.4138 significa que el 41.38 % de la variabilidad en los datos de prueba es explicada por el modelo. Esto indica un ajuste moderado, Un valor cercano a 1 sería ideal, ya que así el modelo explicaría casi toda la variabilidad en los datos
- **R-squared ajustado:** El R-squared ajustado de 0.4127 es ligeramente menor que el R-squared normal. Es decir, teniendo en cuenta el número de predictores, el modelo todavía explica el 41.27 % de la variabilidad en los datos de prueba. (Cosio, 2021)
- **MAE:** Un MAE de 2574355.08 indica que, en promedio, las predicciones del modelo difieren de los valores reales en aproximadamente 2574355.08 unidades. El valor obtenido es bastante alto, lo que señala una desviación significativa en las predicciones.

4 Conclusiones y recomendaciones

De acuerdo con las métricas evaluadas, los modelos de regresión polinomial (Ridge) y regresión polinomial simple muestran un rendimiento similar, con un R^2 alrededor de 0.47, lo que sugiere que estos modelos explican aproximadamente el 47 % de la variabilidad en los datos. Sin embargo, todos los modelos presentan valores de RMSE y MAE relativamente altos, indicando que las predicciones tienen una desviación significativa respecto a los valores reales.

Se recomienda realizar un análisis de residuos para entender mejor dónde están fallando los modelos actuales y ajustar en consecuencia. Esto puede revelar patrones o tendencias que los modelos actuales no están capturando.

Dado que los modelos lineales y polinomiales no están proporcionando predicciones precisas, podría ser beneficioso explorar modelos no lineales más complejos como árboles de decisión, bosques aleatorios, y métodos de boosting (e.g., XGBoost, LightGBM).

5 Link del repertorio de GitHub:

https://github.com/sofiabocker/proyecto_ca0305_

Referencias

B12admark. (2020). *Qué son regresión y clasificación en machine learning.* Recuperado el 2 de mayo de 2024 de https://agenciab12.mx/noticia/que-son-regresion-clasificacion-machine-learning?utm_source=social&utm_medium=facebook&utm_campaign=ia.

Basysyar, F. M., y Dwilestari, G. (2022). Enhanced cybersecurity using artificial intelligence and machine learning. *Advancements in Technology, Artificial Intelligence, and Machine Learning*, *1*(1), 25-37. Descargado de https://library.acadlore.com/ATAIML/2022/1/1/ATAIML_01.01_03.pdf

Burrueco, D. (2022). *Regresión lasso.* Descargado de https://interactivechaos.com/es/manual/tutorial-de-machine-learning/regresion-lasso

Cosio, N. A. L. (2021). *Métricas en regresión.* Descargado de https://medium.com/@nicolasarrioja/m%C3%A9tricas-en-regresi%C3%B3n-5e5d4259430b

DELTA, I. (2020). *Regresión 2. regresión polinomial y regularización.* Recuperado el 1 de julio de 2024 de https://iadelta.com/inteligencia-artificial/regresion/polinomial-y-regularizacion/.

Fernández Batalla, O. (2021). *Extracción de datos y modelo predictor para el precio de alquiler de viviendas de barcelona* (Tesis de Master no publicada). Universitat de Barcelona.

González, D. H. (2008). *Cconceptos básicos de métricas.* http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/gonzalez_d_h/capitulo2.pdf.

Grajales Alzate, Y. V. (2019). *Modelo de predicción de precios de viviendas en el municipio de rionegro para apoyar la toma de decisiones de compra y venta de propiedad raíz* (Tesis de Master no publicada). Escuela de Ingenierías.

IBM. (2021). *Regresión lineal de elastic net.* Recuperado el 1 de julio de 2024 de https://www.ibm.com/docs/es/spss-statistics/saas?topic=features-linear-elastic-net-regression.

RStudio. (s.f.). *Estadísticas de muestreo y entrenamiento de modelos en r.* RPubs. Descargado de https://rpubs.com/joralex0826/749750

Saavedra, J. A. (2023). *Regresión lineal: teoría y ejemplos.* Recuperado el 1 de julio de 2024 de https://ebac.mx/blog/regreson-lineal#:~:text=La%20regresi%C3%B3n%20lineal%20es%20un,utilizado%20en%20el%20aprendizaje%20autom%C3%A1tico.

Sandoval Serrano, L. J., y et al. (2018). Algoritmos de aprendizaje automático para análisis y predicción de datos. *Revista Tecnológica; no. 11*.

Thamarai, M., y Malarvizhi, S. P. (2020). House price prediction modeling using machine learning. *Journal of Real Estate Data Science*, *1*(1), 6.

Vitalflux. (s.f.). *Mse vs rmse vs mae vs mape vs r-squared: When to use which measure?* https://vitalflux.com/mse-vs-rmse-vs-mae-vs-mape-vs-r-squared-when-to-use/.