Parcial 2

Sebastián Ramírez Escobar, ✉ sramireze1@eafit.edu.co

Profesor: Pablo Andres Saldarriaga Aristizabal

-

Universidad EAFIT
Estadística no Paramétrica en Ciencias de los datos
Medellín
2024

TABLA DE CONTENIDOS

LISTA DE TABLAS

LISTA DE FIGURAS

I. PUNTO 1

I-A. Parte 1

La integral que necesitamos evaluar es:

$$\int_0^2 \int_0^2 K(x+y)\,dx\,dy$$

Integrando con respecto a x:

$$\int_0^2 K(x+y)\,dx = K\left[\frac{x^2}{2} + xy\right]_0^2 = K(2+2y)$$

Integrando con respecto a y:

$$\int_0^2 K(2+2y)\,dy = K\left[2y + y^2\right]_0^2 = 8K$$

Para que la función sea una densidad de probabilidad, la integral debe ser igual a 1:

$$8K = 1 \quad \Rightarrow \quad K = \frac{1}{8}$$

El valor correcto de K que hace que $f(x,y) = K(x+y)$ sea una función de densidad de probabilidad válida es $\frac{1}{8}$.

I-B. Parte 2

$$f(x,y) = \frac{1}{8}(x+y) \quad \text{para} \quad 0 < x < 2 \text{ y } 0 < y < 2$$

Para obtener la función de densidad marginal $f(x)$, integramos la función de densidad conjunta $f(x,y)$ con respecto a y sobre el intervalo permitido:

$$f(x) = \frac{1}{8}\left(\int_0^2 x\,dy + \int_0^2 y\,dy\right)$$

$$f(x) = \frac{1}{8}\left(x \cdot [y]_0^2 + \frac{y^2}{2}\Big|_0^2\right)$$

Evaluando los términos de la integral:

$$f(x) = \frac{1}{8}\left(2x + [y^2/2]_0^2\right)$$

$$f(x) = \frac{1}{8}\left(2x + 2\right)$$

$$= \frac{1}{4}(x+1)$$

I-C. Parte 3

La función de densidad conjunta es dada por:

$$f(x,y) = \frac{1}{8}(x+y)$$

La función de densidad marginal:

$$f(x) = \frac{1}{4}(x+1)$$

La función de densidad condicional:

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{\frac{1}{8}(x+y)}{\frac{1}{4}(x+1)} = \frac{x+y}{2(x+1)}$$

Usando la fórmula de la esperanza condicional:

$$E(Y|X) = \int y f(y|x)\, dy$$

$$E(Y|X=x) = \int_0^2 y \frac{x+y}{2(x+1)}\, dy$$

Desarrollando la integral:

$$E(Y|X=x) = \frac{1}{2(x+1)} \int_0^2 y(x+y)\, dy$$

$$= \frac{1}{2(x+1)} \left(\int_0^2 xy\, dy + \int_0^2 y^2\, dy \right)$$

$$= \frac{1}{2(x+1)} \left(2x + \frac{8}{3} \right) = \frac{2x + \frac{8}{3}}{2(x+1)}$$

$$E(Y|X=x) = \frac{x + \frac{4}{3}}{x+1}$$

I-D. Simulación de los datos

Se realizó una simulación de 1000 valores de Y con desviación de 0.1 y ruido blanco gaussiano de $E(Y|X = x) = \frac{x+\frac{4}{3}}{x+1}$. Se obtiene la siguiente gráfica.

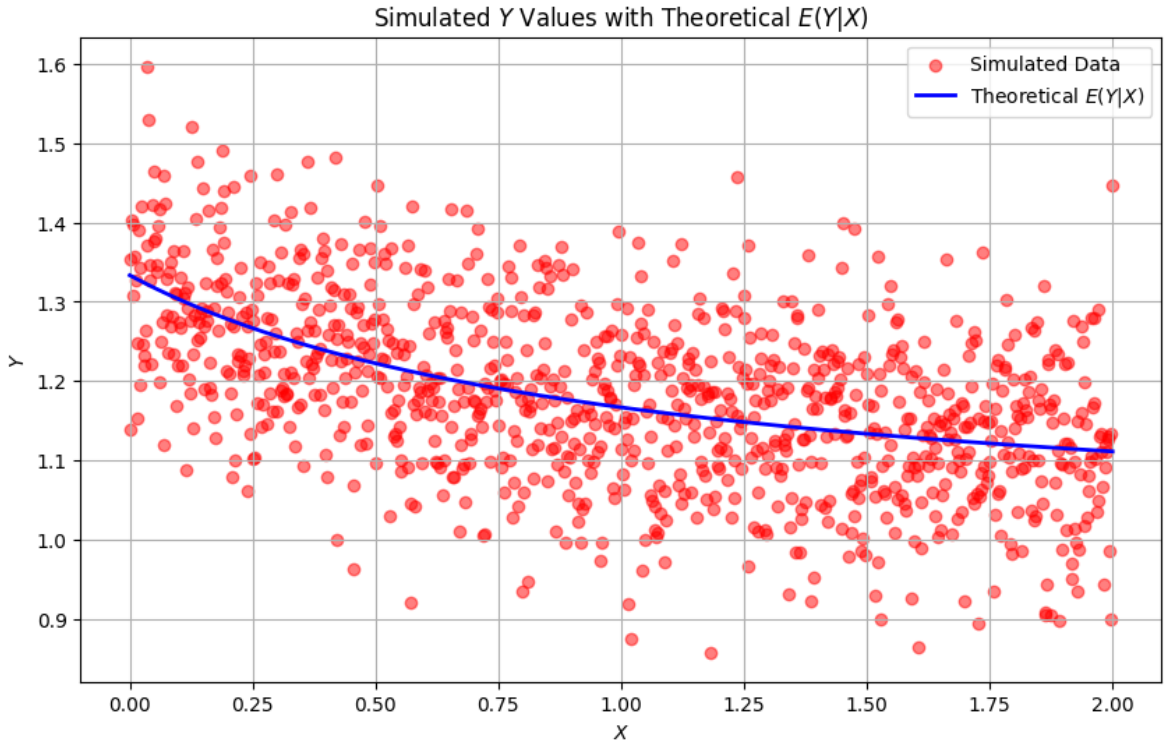

FIG. 1: Valores simulados con E(Y—X)

I-E. Regresiones

Se realizaron las regresiones paramétrica (Lineal) y no paramétrica (Kernel), se evaluaron con las métricas R^2 y RMSE.

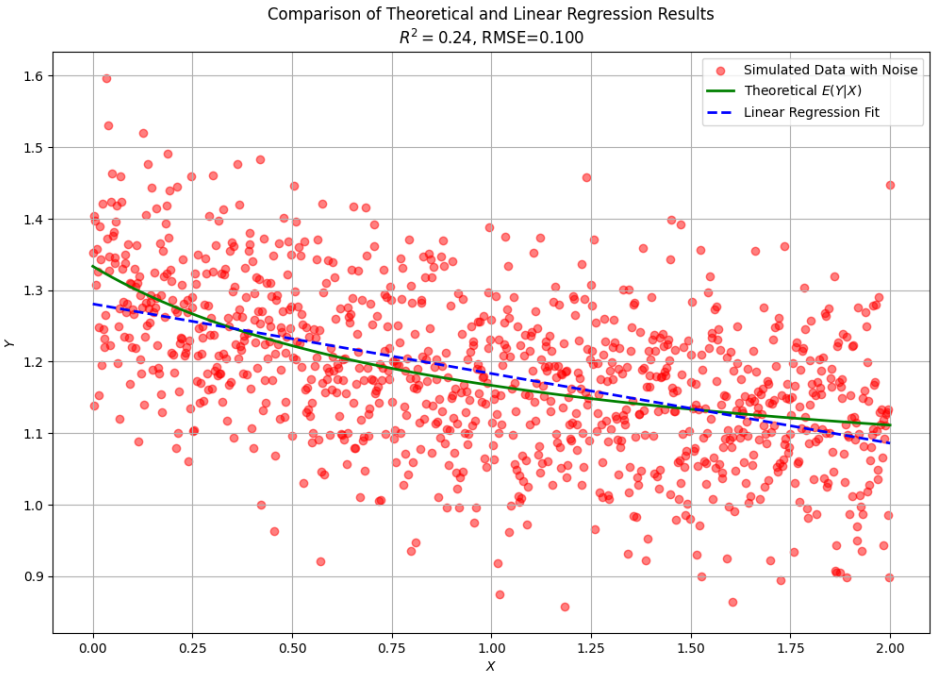

FIG. 2: Regresión lineal y Modelo Teórico

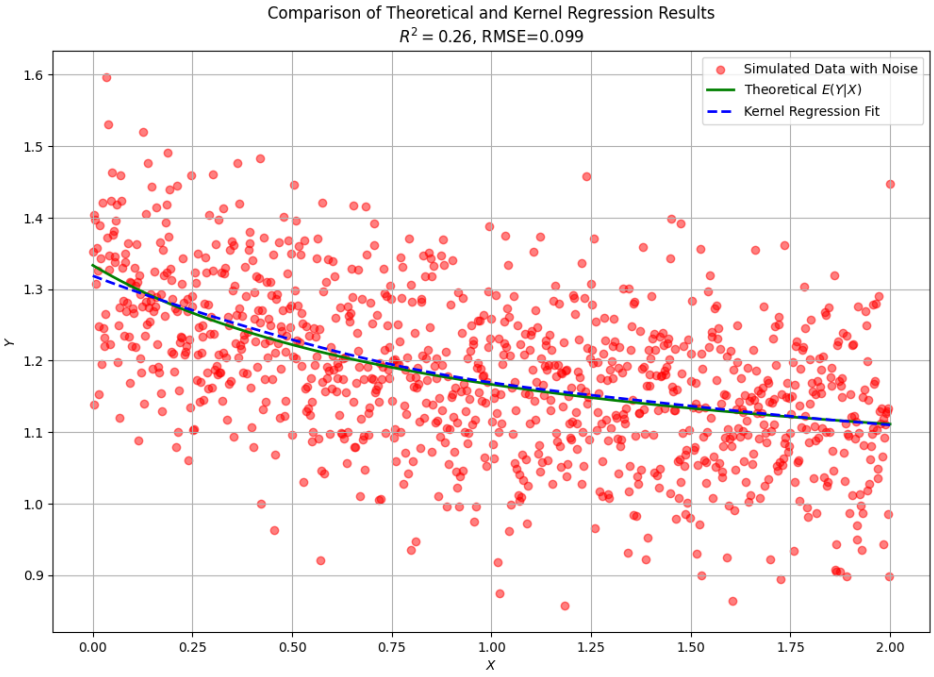

FIG. 3: Regresión Kernel y Modelo Teórico

I-E1) Análisis de Resultados: Aunque ambos valores de R^2 son relativamente bajos, lo que sugiere que ninguno de los modelos capta completamente la variabilidad de los datos, el modelo de regresión kernel muestra un leve mejoramiento en comparación con el modelo lineal.

Ambos modelos tienen un RMSE casi idéntico, indicando una precisión similar en términos del error promedio en las predicciones.

El modelo de regresión kernel es más adecuado para estos datos.

II. Punto 2

II-A. Inciso a - KNN Mahalanobis

Se utilizó la inversa matriz de covarianza en la clase de la librería KNN Regressor de sklearn, utilizando la distancia de mahalanobis como distance_metric. El desempeño en los datos reales vs predichos en ajuste a la línea de predicción ideal:

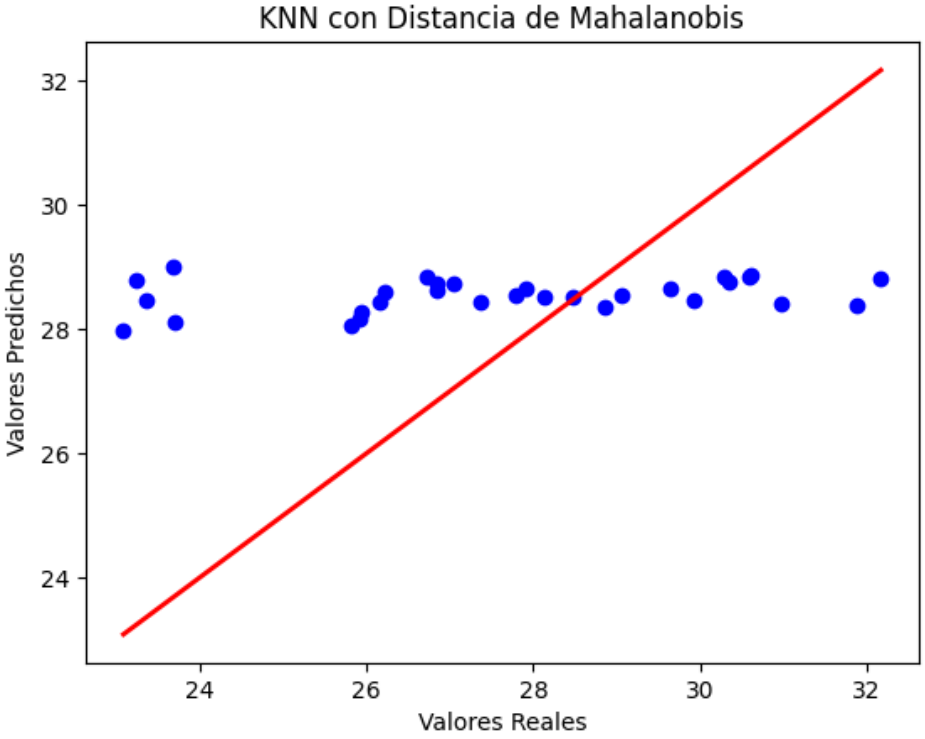

FIG. 4: Ajuste datos reales vs predichos KNN Mahalanobis

Métrica **Best K = 43**	**Train**	**Test**
Mean Squared Error (MSE)	9.7058	7.8729

TABLA. I: Resultados del KNN Regressor con Distancia de Mahalanobis

II-A1) Análisis de resultados: El modelo KNN regressor con distancia de Mahalanobis y el mejor valor de K=43 muestra un rendimiento insatisfactorio: un alto error cuadrático medio (MSE) tanto en el conjunto de entrenamiento como en el de prueba. El gráfico de dispersión refuerza esta conclusión al mostrar una dispersión significativa de las predicciones alrededor de la línea ideal, sugiriendo una pobre correlación entre los valores reales y predichos.

II-B. Inciso b - KNN Mahalanobis Robusto, Regresiones: Lineal y Robusta RLM con constante

Se realizan 3 regresiones en el conjunto de datos con los siguientes modelos:

- KNN Mahalanobis Robusto: Recorte de matriz de covarianza con Ledoit-Wolf
- Regresión Lineal
- Regresión Lineal Robusta con Constante (RLM)

II-B1) KNN Mahalanobis Robusto: Recorte de matriz de covarianza con Ledoit-Wolf: La técnica de Ledoit-Wolf proporciona un estimador de la matriz de covarianza que es una combinación ponderada entre la matriz de covarianza empírica y la matriz de identidad Ledoit y Wolf, 2004, lo que reduce el sesgo y la varianza del estimador en presencia de datos ruidosos o pequeños tamaños de muestra. Esta regularización controla el problema de matrices de covarianza mal condicionadas y hace que el cálculo de la distancia de Mahalanobis sea más estable. Se utiliza la función creada en el inciso a, modificando la inversa de la matriz de covarianza.

Métrica **Best K = 4**	**Train**	**Test**
Mean Squared Error (MSE)	2.0325	1.0436

TABLA. II: Resultados del KNN Regressor con Distancia de Mahalanobis Robusta (Ledoit-Wolf)

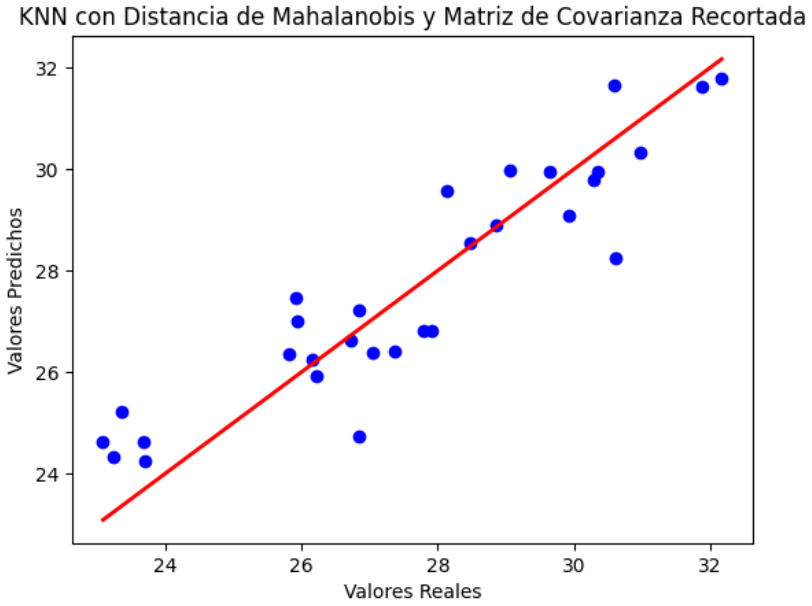

FIG. 5: Ajuste datos reales vs predichos KNN Mahalanobis Robusto

Con un valor de K=4, el modelo KNN Mahalanobis robusto alcanza un error cuadrático medio (MSE) de 2.1861 en el conjunto de entrenamiento y 1.0575 en el conjunto de prueba, indicando una precisión alta en las predicciones y su relación entre resultados de train y test es coherente.

II-B2) Regresión lineal: Aplicando la regresión se obtienen los siguientes resultados

Métrica	**Train**	**Test**
Mean Squared Error (MSE)	0.3097	1.3495

TABLA. III: Resultados de la Regresión Lineal

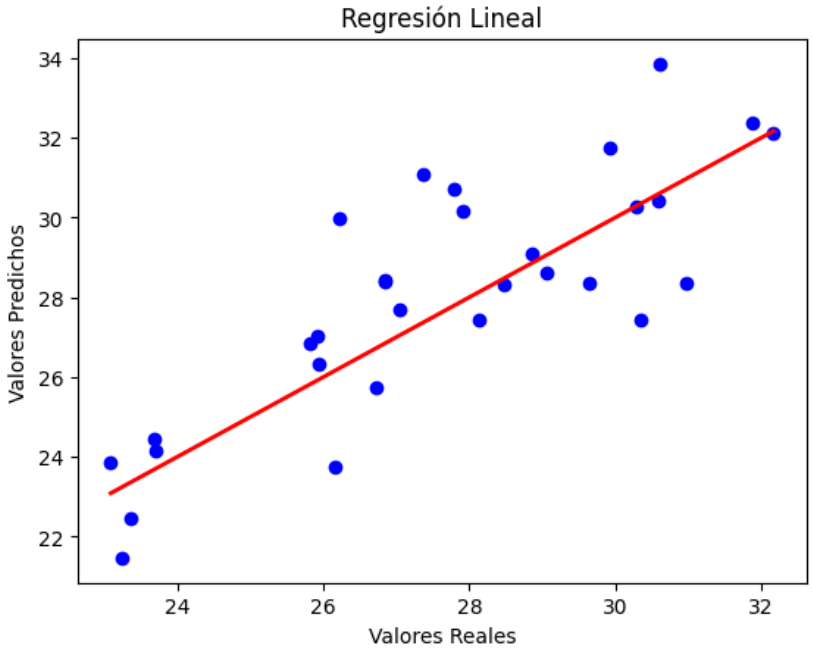

FIG. 6: Ajuste datos reales vs predichos Regresión Lineal

La tabla de resultados indica un error cuadrático medio (MSE) de 0.3 en el conjunto de entrenamiento y 1.3 en el de prueba, indicando que el modelo se ajusta bien a los datos de entrenamiento pero tiene un desempeño significativamente peor en los datos de prueba, lo que parece ser un sobreajuste del modelo a los datos de entrenamiento.

II-B3) Regresión lineal robusta con coeficiente (RLM): A diferencia de la regresión lineal ordinaria que minimiza la suma de los errores al cuadrado (MSE), la RLM utiliza métodos de estimación robusta, como M-estimadores, que minimizan una función de pérdida menos sensible a valores atípicos. Esto se logra ponderando los errores de acuerdo a su magnitud, de manera que los outliers tienen menos influencia en el ajuste del modelo.

Métrica	Train	Test
Mean Squared Error (MSE)	0.3121	1.3407

TABLA. IV: Resultados de la Regresión Lineal Robusta (RLM)

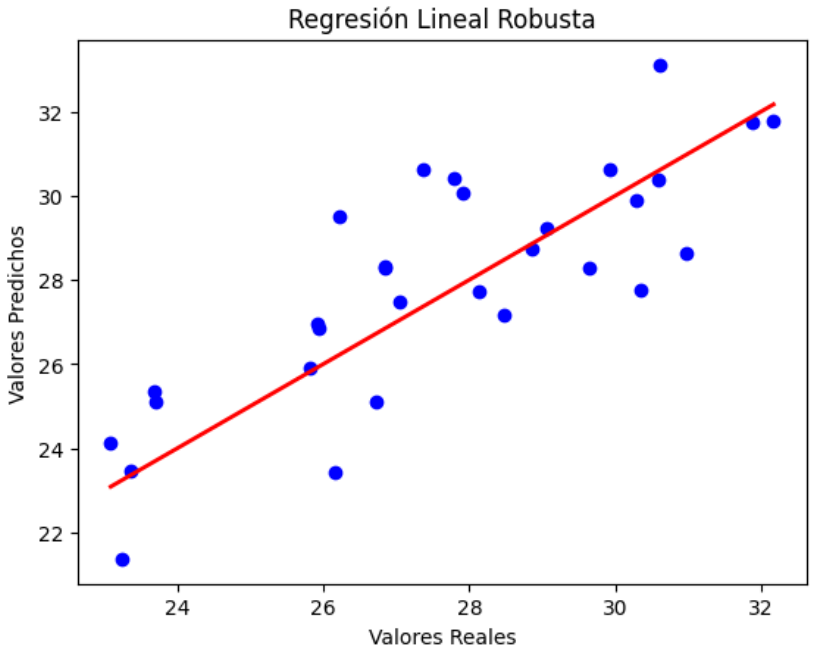

FIG. 7: Ajuste datos reales vs predichos Regresión Lineal Robusta

El modelo robusto indica una capacidad de generalización razonable. Un poco mejor que el modelo de regresión lineal en los datos de prueba. Estos resultados sugieren un poco de sobre ajuste a los datos de entrenamiento, sin embargo, en prueba tiene un rendimiento aceptable.

Modelo	**Métrica**	**Train**	**Test**
Regresión Lineal	MSE	0.3097	1.3495
Regresión Lineal Robusta (RLM)	MSE	0.3121	1.3407
KNN con Mahalanobis Robusta (Ledoit-Wolf)	MSE	2.0325	1.0436

TABLA. V: Comparación de Resultados de los Modelos

II-B4) Análisis de resultados: Al comparar los tres modelos, el KNN con distancia de Mahalanobis robusta (Ledoit-Wolf) presenta el mejor ajuste en el conjunto de validación e indicando una excelente capacidad de generalización. En el conjunto de entrenamiento, la regresión lineal muestra el mejor ajuste, pero se observa un significativo sobreajuste, dado el aumento del MSE en el conjunto de prueba. La regresión lineal robusta (RLM) ofrece un equilibrio intermedio un poco menos sobreajustado, con un ajuste levemente mejor tanto en entrenamiento como en prueba, mostrando menor capacidad predictiva en comparación con el KNN robusto en validación. Estos resultados sugieren que el KNN robusto es más efectivo en capturar las relaciones subyacentes en datos nuevos, mientras que la regresión lineal simple es propensa al sobreajuste, y la RLM mejora un poco con respecto a la regresión simple.

La regresión lineal puede estar sobreajustandose a los datos mal condicionados en el conjunto de entrenamiento, capturando ruido y no la distribución subyacente de los datos. La RLM puede tener el mismo problema, ya que el conjunto de datos tiene ruido agregado. El KNN mahalanobis robusto no tiene problema con los datos mal condicionados ya que el uso de la matriz de covarianza recortada mediante Ledoit-Wolf proporciona estimaciones más estables y menos influenciadas por ruido y otros factores como outliers, lo que mejora la capacidad del modelo para generalizar en el conjunto de prueba.

II-C. Inciso c

Se agregaron outliers cell wise al dataset, correspondientes al 15 % de los datos totales. Se realiza un nuevo entrenamiento de los modelos en el dataset contaminado de entrenamiento y posteriormente se predice en el total del dataset contaminado de prueba. Finalmente se procede a evaluar con el error absoluto mediano, que mide la mediana de los errores absolutos entre las predicciones y los valores reales. A diferencia del error absoluto medio (MAE), que promedia los errores absolutos, el MedAE toma la mediana, proporcionando así una medida central del error robusta a outliers.

$$\text{MedAE} = \text{mediana}\left(|y_i - \hat{y}_i|\right)$$

Modelo	**MedAE (train)**	**MedAE (test)**
KNN Mahalanobis robusto (k=64)	1.0352	0.5579
Regresión Lineal	1.1176	1.7402
Regresión Lineal Robusta (RLM)	1.0767	1.7232

TABLA. VI: Resultados de MedAE para diferentes modelos con y sin outliers

Los resultados indican que el KNN robusto tiene un desempeño significativamente mejor al manejar outliers en comparación con los modelos de regresión lineal.

La razón detrás de estos resultados puede estar en la naturaleza de los modelos. La regresión lineal tradicional y robusta son sensibles a los outliers, ya que intentan ajustar una línea recta a todos los puntos de datos, incluyendo aquellos que son atípicos.

Los resultados también sugieren que al agregar una cantidad significativa de outliers se modifica la distribución normal inicial perjudicando los modelos de regresión lineal, en específico al modelo que no es robusto, aunque no difieren mucho en la métrica puesto que la regresión lineal robusta (RLM) es diseñada para mitigar el efecto de los outliers, sigue siendo afectada significativamente cuando la proporción de outliers es considerablemente alta. Caso contrario al modelo no paramétrico y robusto (KNN Mahalanobis robusto). Este modelo no paramétrico clasifica los puntos de datos basándose en la proximidad a sus vecinos más cercanos. La distancia de Mahalanobis considera la estructura de la covarianza de los datos, lo que permite medir la distancia entre puntos en un espacio transformado donde las dimensiones están escaladas de acuerdo con su variabilidad y correlaciones. La matriz de covarianza recortada mediante Ledoit-Wolf ajusta las estimaciones de covarianza para ser más estables y menos influenciadas por outliers.

III. PUNTO 3

III-A. Test de normalidad Shapiro-Wilk

Realizando un gráfico de las distribuciones de las regiones, se obtiene:

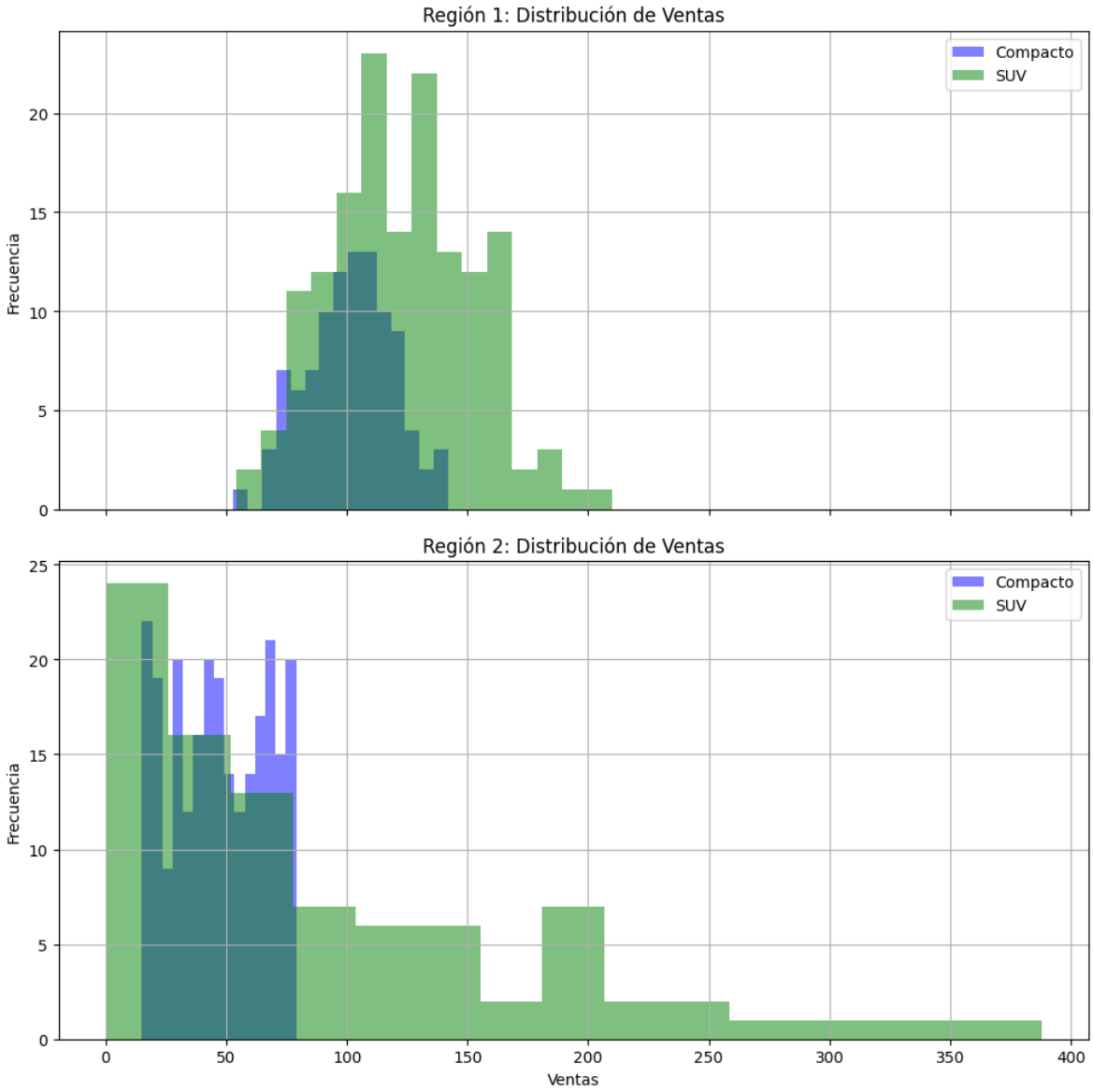

FIG. 8: Histogramas de distribuciones de ventas para las dos regiones

Las distribuciones de las ventas de la región 1 parecen seguir una distribución normal. Se aplica el test de Shapiro Wilk para verificar si los datos de ventas de la región 1 y 2 vienen de una distribución normal.

- **H0:** La muestra de datos sigue una distribución normal.
- **H1:** La muestra de datos no sigue una distribución normal.

TABLA. VII: Resultados del test de Shapiro-Wilk para las ventas de automóviles

Variable	**Estadístico de Prueba**	**Valor p**
Ventas Compacto (Región 1)	0.992	0.837
Ventas Compacto (Región 2)	0.949	$1{,}05 \times 10^{-7}$
Ventas SUV (Región 1)	0.994	0.808
Ventas SUV (Región 2)	0.857	$7{,}32 \times 10^{-8}$

Para un nivel de significancia de 0.05, se concluye que las ventas para la región 1 siguen una distribución normal, mientras que para la región 2 no hay evidencia significativa de que los datos vengan de una distribución normal.

Teniendo en cuenta el test anterior para la región 1 se utilizará el test de t student, para la región 2 se utilizará el test no paramétrico de Mann Whitney-U.

III-B. Inciso a

III-B1) Prueba 1 - 'two sided':

- **H0:** No existe diferencia significativa en las medias de las ventas de los dos modelos de automóviles en las dos regiones.
- **H1:** Existe una diferencia significativa en las medias de las ventas de los dos modelos de automóviles en las dos regiones.

III-B2) Prueba 2 - 'less': **Hipótesis para Región 1:**

- **H0:** Las ventas del modelo compacto son mayores o iguales a las del modelo SUV.
- **H1:** Las ventas del modelo SUV son mayores que las del modelo compacto.

Hipótesis para Región 2:

- **H0:** Las ventas del modelo compacto son mayores o iguales a las del modelo SUV.
- **H1:** Las ventas del modelo SUV son mayores que las del modelo compacto.

El nivel de significancia es de 0.05 para las dos pruebas

III-C. Inciso b

Los resultados de la prueba de Mann-Whitney U para las dos regiones son los siguientes:

III-C1) Resultados Prueba 1:

- **Región 1:**
- Estadístico de T: -6.4037
- Valor p: $7{,}53 \times 10^{-9}$

- **Región 2:**
- Estadístico de Mann-Whitney U: 9001.0
- Valor p: 0.0049

Para ambas regiones se rechaza la hipótesis nula y concluimos que existe una diferencia significativa en las ventas de los dos modelos de automóviles en ambas regiones.

III-C2) Resultados Prueba 2:

- **Región 1:**
- Estadístico de T: -6.4037
- Valor p: $3{,}77 \times 10^{-9}$

- **Región 2:**
- Estadístico de Mann-Whitney U: 9001.0
- Valor p: 0.0025

Para ambas regiones se rechaza la hipótesis nula y se concluye que las ventas del modelo SUV son significativamente mayores que las del modelo compacto.

IV. Punto 4

Para la encuesta, se plantea el test de wilcoxon para datos pareados con un nivel de significancia de 0.05.

IV-A. Test de Hipótesis

- **H0:** No hay diferencia significativa en el número de viajes fuera de la ciudad entre 2023 y 2022.
- **H1:** Hay una tendencia a realizar más viajes fuera de la ciudad en 2023 que en 2022.

IV-B. Resultados del test

- Estadístico de la prueba: 342.0
- Valor p: 0.00011

Conclusión: Rechazamos la hipótesis nula. Hay evidencia suficiente para decir que hay una tendencia a viajar más en 2023 que en 2022.

REFERENCIAS

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365-411. https://doi.org/ https://doi.org/10.1016/S0047-259X(03)00096-4