Campus Irapuato-Salamanca | División de Ingenierías

Regresión Lineal y K-NN: Introducción al machine learning

Cervantes Martínez Rodrigo Fabián

26/02/2024

Resumen

En este informe se plantea el uso de dos técnicas de modelado como un primer acercamiento a la estadística inferencial, la regresión lineal simple y el k-vecinos más cercanos (K-NN) son técnicas ideales, pues su simplicidad permite una mayor comprensión, esto con la finalidad de explorar y analizar conjuntos de datos a partir de las estadísticas que estos proporcionan.

Palabras clave

Estadística inferencial, modelado, regresión lineal simple, K-NN.

1. Introducción

El modelado desempeña un papel esencial en la ciencia de datos y el aprendizaje automático, ya que forma parte del proceso mediante el cual se crea una representación simplificada y estructurada de un fenómeno o sistema del mundo real. Bajo el contexto de la regresión lineal y el k-vecinos más cercanos (KNN), la modelización implica la construcción de fórmulas matemáticas o el encuentro de patrones que describen la relación entre variables.
En esta exploración, profundizaremos en cómo la modelización a través de los métodos ya dichos, proporciona herramientas valiosas para entender y predecir datos en diversos contextos, ofreciendo así una visión detallada de la construcción de modelos y su aplicación práctica en el análisis de datos.

2. Trabajos relacionados

La regresión lineal y el KNN son métodos ampliamente estudiados y enfocados a diferentes campos, siendo técnicas cruciales para la predicción de datos. A continuación, se presenta una revisión de trabajos relacionados que han explorado y ampliado el conocimiento en estas áreas:

1. **An efficient instance selection algorithm for k nearest neighbor regression:** [6]
   Este trabajo se centra en la selección de instancias para la regresión del vecino más cercano k, eliminando las instancias atípicas, ordenando los datos y eliminando una por una de las instancias que no afecten la regresión lineal.

2. **Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock:** [1]
   En este estudio comparativo se plantea que el método Mínimos Cuadrados Ordinarios (OLS) suele ser un modelo con mejores resultados en predicción que KNN, aunque por muy poco. Además se demuestra que con cierto tipo de distancias, KNN tiende al sobreajuste, lo que hace del modelo un poco más complejo de llevar a cabo.

3. **Application of Improved Linear Regression Algorithm in Business Behavior Analysis:** [5]
   Esta investigación propone que la regresión lineal es un método bastante efectivo y útil en el área de negocios, pues ayuda a tomar decisiones con un trasfondo sólido, lo que mejora el estudio de mercado y las estrategias de venta.

Estos trabajos muestran la diversidad de aplicaciones y enfoques que rodean la regresión lineal y el KNN, proporcionando una base sólida para futuras investigaciones.

3. Materiales y métodos

Materiales:

- **Datos de entrenamiento y prueba:**
  Conjuntos de datos proporcionados para el análisis, una relación de variable dependiente $\boldsymbol{Y}$ con una variable independiente $\boldsymbol{X}$.

- **Herramientas de programación:**
  Entorno de Google Colab, Python con bibliotecas cómo Numpy, Pandas y Scikit-Learn para la implementación de los modelos de regresión lineal, matplotlib visualización de datos, y stats para pruebas de normalidad.

Métodos:

- **Regresión lineal:**
  La regresión lineal simple es un modelo estadístico que busca establecer una relación lineal entre una variable dependiente $\boldsymbol{Y}$ y una variable independiente $\boldsymbol{X}$. El método más usado para realizar esto es el de Mínimos Cuadrados Ordinarios (OLS) cuyo objetivo es encontrar los coeficientes que minimizan la suma de los cuadrados de los residuos, es decir, las diferencias entre los valores obsservados y los valores predichos por el modelo.[3]

  La ecuación de regresión lineal simple con el método OLS se expresa como:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  donde:

  - Y es la variable dependiente que queremos predecir.
  - X es la variable independiente.
  - β_0 es la ordenada al origen o intercepto.
  - β_1 es la pendiente de la recta de regresión.
  - ϵ es el término de error.

  El objetivo del método OLS es encontrar los valores de β_0 y β_1 que minimizan la función de costo, que es la suma de los cuadrados de los residuos. La función de costo se define como:

$$\text{Costo} = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

  Donde $\boldsymbol{n}$ es el número de observaciones en el conjunto de datos.
  Los valores óptimos de β_0 y β_1 que minimizan esta función se obtienen calculando derivadas e igualándolas a cero. La solución cerrada para los coeficientes es:

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

  donde $\bar{X}$ y $\bar{Y}$ son las medias de las variables $\boldsymbol{X}$ e $\boldsymbol{Y}$, respectivamente. Una vez que se encuentran estos coeficientes, se puede utilizar la ecuación de regresión lineal para hacer predicciones sobre $\boldsymbol{Y}$ para nuevos valores de $\boldsymbol{X}$.

- **K-Vecinos Más Cercanos (KNN):**
  El algoritmo K-Vecinos más cercanos para regresión es una técnica de aprendizaje

supervisado que utiliza la información de los vecinos más cercanos para predecir valores numéricos en lugar de clasificar instancias.

Dada una muestra de entrenamiento(x_i, y_i), donde x_i es un vector decaracterísticas y y_i es la variable dependiente (valor a predecir), el modelo KNN estima $\hat{y}$, el valor predicho para una nueva entrada x_{new}, utilizando los k vecinos más cercanos:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} w_i \cdot y_i$$

donde:

- w_i es el peso asociado al i-ésimo vecino, calculado mediante la inversa de la distancia entre x_{new} y x_i:
$$w_i = \frac{1}{d_i}$$
- d_i es la distancia entre x_{new} y x_i, que puede ser calculada usando alguna métrica de distancia, Euclidiana, Manhattan o Minkowski.

En resumen, KNN para regresión lineal promedia los valores de la variable dependiente de los k vecinos más cercanos, ponderados por la inversa de sus distancias al punto de consulta.

- **Distancia de Minkowski:**
La distancia de Minkowski entre dos vectores $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ e $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ se define como:
$$D(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$
Donde $\|\mathbf{x} - \mathbf{y}\|_p$ representa la norma p de la diferencia entre $\mathbf{x}$ e $\mathbf{y}$.

4. Experimentos

La razón principal de los experimentos realizados es conocer cómo funcionan ambos métodos para regresión lineal y poder realizar un contraste entre su funcionamiento y eficiencia. Para este análisis se proporcionaron 3 conjuntos de datos:

- **Dataset 1:**
Tiene solo 2 columnas y 500 instancias, la variable independiente ('cero') y la dependiente ('uno'), en este caso no hay mayor información, parece ser una mera relación de ejes X e Y.

- **Dataset 2:**
Nuevamente hay 2 columnas, pero 506 intancias, el objetivo que es el precio de una casa, y la variable independiente, la tasa de crimen en esa región.

- **Dataset 3:**
  Este conjunto de datos tiene 26 características y 205 instancias que describen automóviles, entre ellas, se encuentra la variable a predecir, el precio.

Para estos 3 casos se realizó primeramente una visualización de su comportamiento, pruebas de normalidad debido a que es un requisito primordial para realizar una regresión lineal, cálculo de correlación entre las variables y un análisis de características para asegurar que no existan datos faltantes o variables categóricas que tratar.
La partición de los conjuntos de datos fue 70 % entrenamiento y 30 % prueba, esto debido a las dimensiones de los datasets, al tener un número considerable de instancias, podemos dejar un porcentaje un tanto alto para realizar las pruebas y un poco más de la mitad para realizar el entrenamiento.
Después de esto, se reescalaron los datos para poder generar los modelos de regresión lineal al cuál también se le encontró un intervalo de confianza y KNN, seleccionando en todos los casos la K óptima (4) así cómo la distancia, que para todos resultó ser la distancia Minkowski. Finalmente se calcularon las métricas para evaluar la efectividad de los modelos generados.

Para el tercer conjunto de datos, primero se realizó una matriz de correlación de las variables a tomar en cuenta para seleccionar las más fuertes con nuestra variable de intéres, que es el precio.

Figura 1: Selección de características

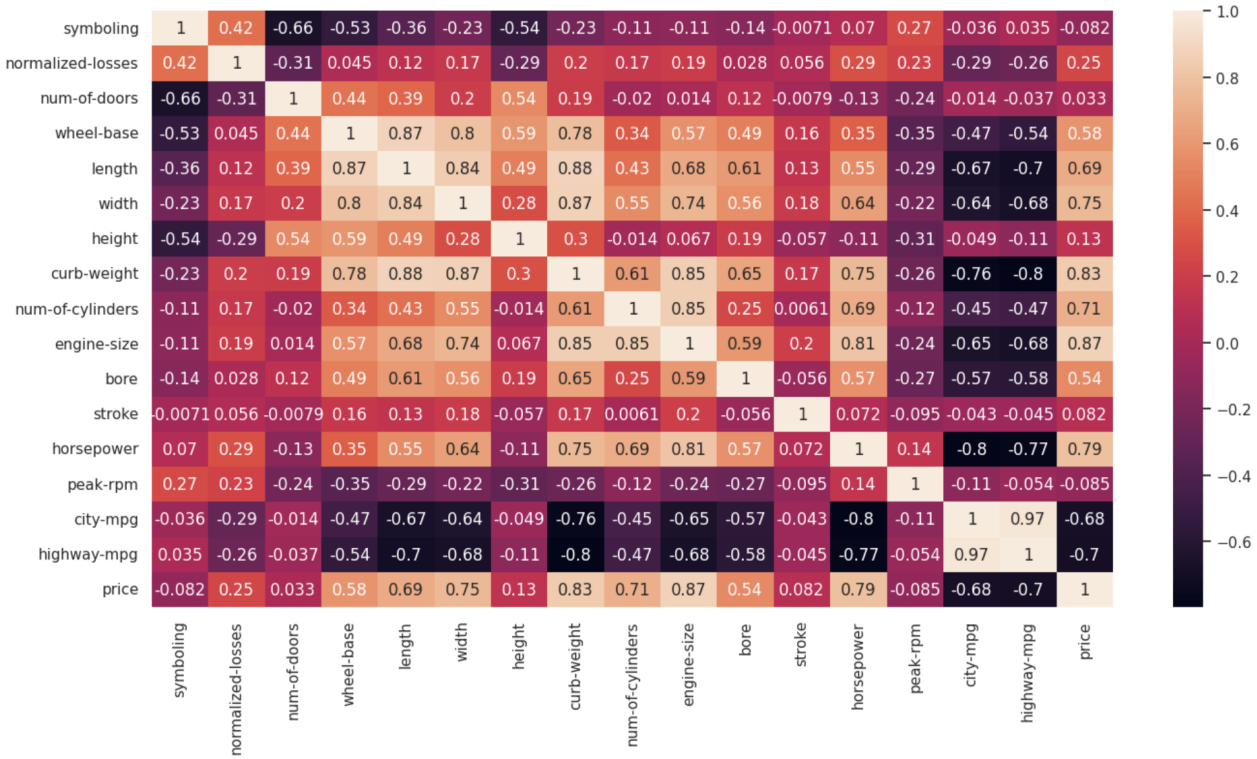
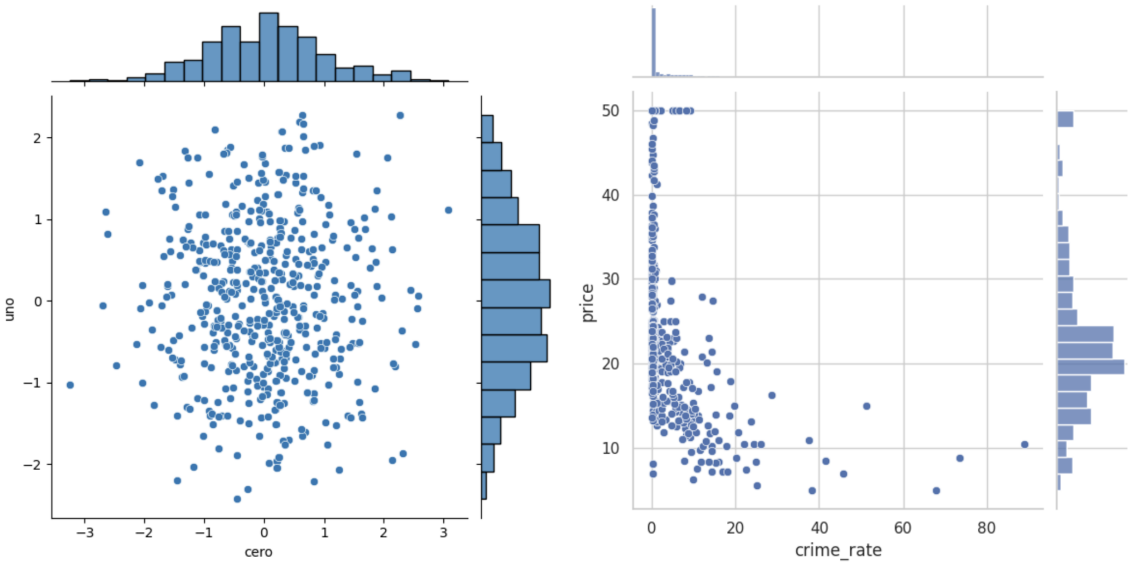
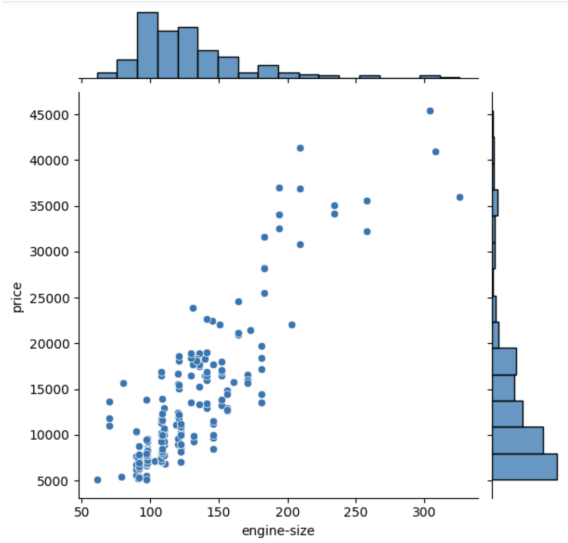


Figura 2: Visualización de datos



(a) Dataset 1

(b) Dataset 2

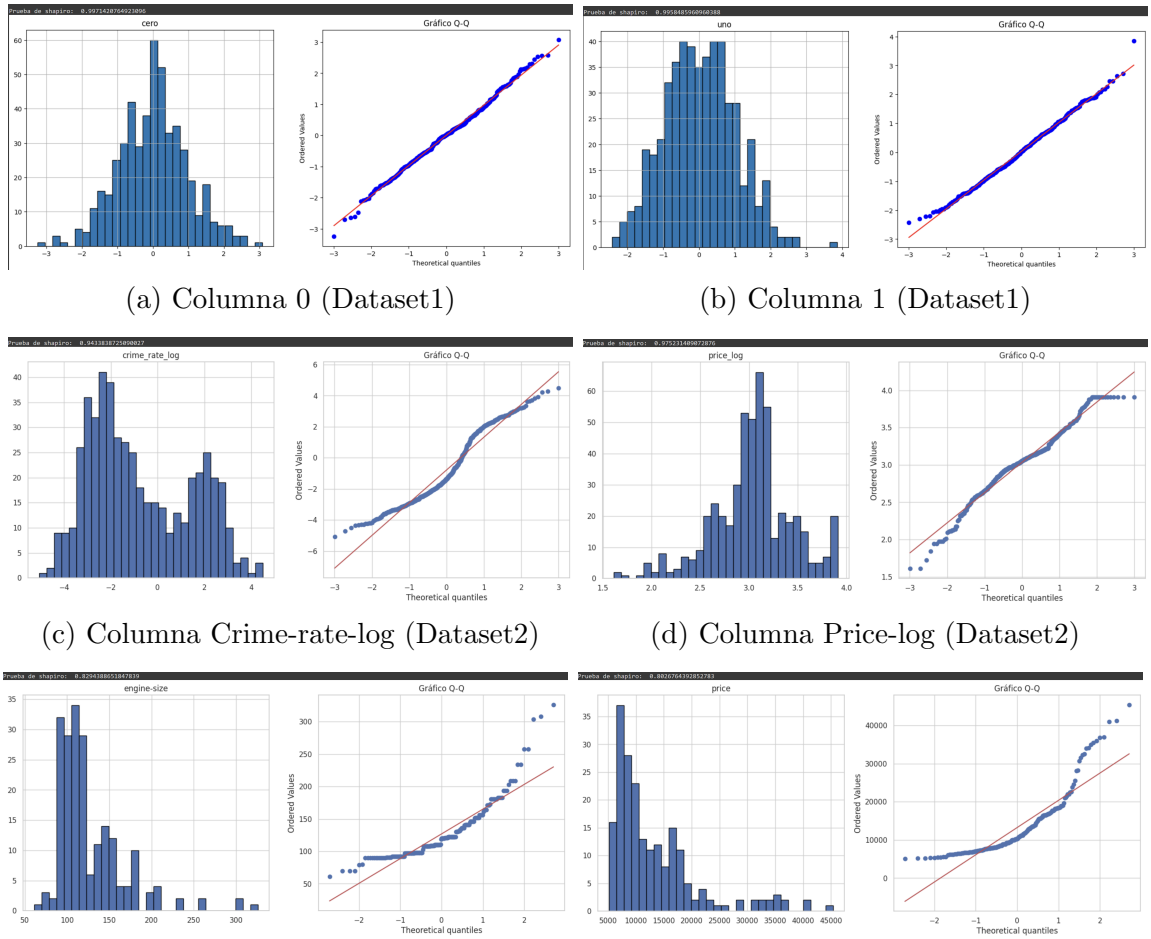


(c) Dataset 3

Según la correlación Pearson[2], en el primer conjunto de datos se tiene una correlación muy débil, en el segundo una fuerte y en el tercero las carecterísticas que pueden predecir mejor el precio, son: 'engine-size', 'curb-weight' y 'horsepower'. Con este primer acercamiento se pudo observar que en el primer conjunto de datos parece que no existe una relación sólida entre las variables, además que ambas columnas se asemejan demasiado a una distribución normal, para el segundo conjunto se puede apreciar que los datos tienen similitud a una función exponencial, lo que supone una necesaria tranformación de datos como parte de su preprocesamiento, y en el tercero se ve claramente una relación fuerte y positiva entre la

variable que escogimos (engine-size) y el objetivo (precio). (Véase Figura 2)

Figura 3: Pruebas de normalidad



(a) Columna 0 (Dataset1)

(b) Columna 1 (Dataset1)

(c) Columna Crime-rate-log (Dataset2)

(d) Columna Price-log (Dataset2)

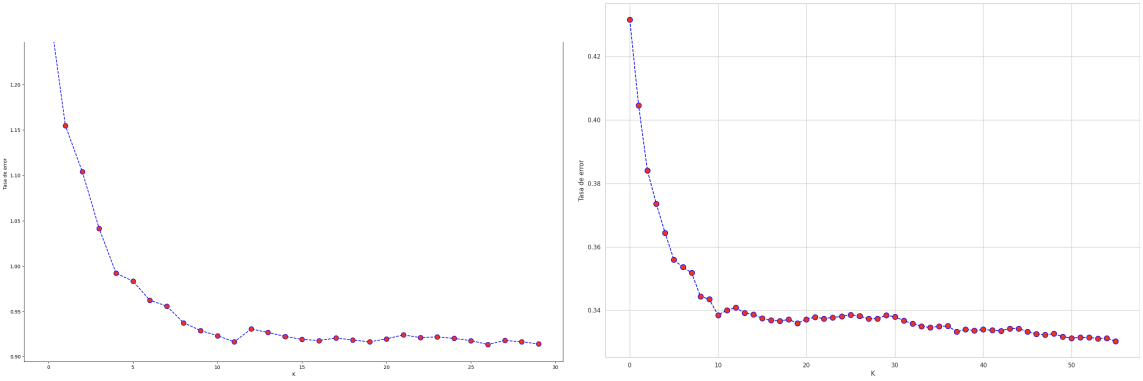
(e) Columna engine-size (Dataset3)

(f) Columna Price (Dataset3)

En el segundo dataset se realizó una transformación logarítimica a ambas columnas para tener un poco más de normalidad en los datos. (Véase Figura 3) Después de realizar las pruebas de normalidad con gráficos cuántil-cuántil y utilizando el método Shapiro-Wilk[4], se puede asegurar que los datos ya siguen o se asemejan bastante a una distribución normal, por lo que se puede continuar con la construcción del modelo y el calculo de los intervalos de confianza.

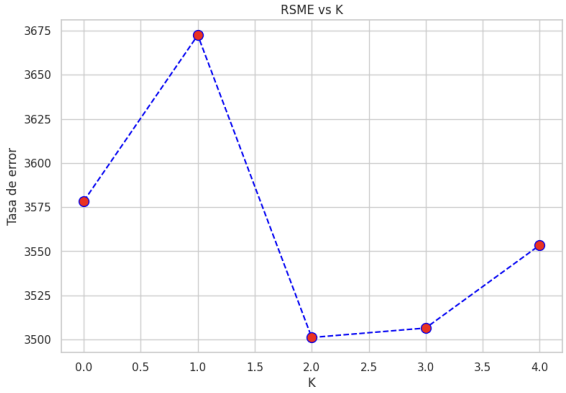
Como se puede apreciar, en todas las métricas de los 3 conjuntos de datos (1, 2, 3), el algoritmo de regresión KNN superó indiscutiblemente a la regresión lineal simple, esto se debe a una de sus grandes diferencias, por una parte, KNN es un método "perezoso", lo que indica que para cada instancia nueva, conserva todos los datos de entrenamiento, situación que en la regresión lineal no ocurre.

Figura 4: K-Óptima



(a) Dataset 1, K=30

(b) Dataset 2, K=54



(c) Dataset 3, K=3

Se eligieron esas 3 diferentes métricas para poder realizar un contraste más amplio, además de estar incluidas las que mejor califican al modelo, cómo R^2 que indica el porcentaje de variabilidad que la regresión lineal puede explicar o MSE que representa el promedio de los cuadrados de las diferencias entre lo que predice el modelo y el valor real.

Se utilizará la métrica Error Medio Cuadrático(RSME), interpretada cómo la desviación estándar del error para los intervalos de confianza de nuestros modelos generados por OLS.(4)

5. Discusión

En la sección de Experimentos, se realizaron diversos análisis y pruebas utilizando conjuntos de datos variados. Se observó que, en general, el algoritmo de regresión KNN superó a la regresión lineal simple en todas las métricas evaluadas. Esto puede atribuirse a la naturaleza "perezosa"de KNN, que conserva todos los datos de entrenamiento para cada nueva instancia, mientras que la regresión lineal sigue un enfoque más simplificado.

En el primer conjunto de datos, donde no parece haber una relación sólida entre las variables, ambos modelos presentaron un rendimiento limitado, pero KNN logró superar

Métrica	RL	K-NN
R^2 (train)	0.0005	0.0399
R^2 (test)	0.0009	0.0238
MSE (train)	1.0019	0.9723
MSE (test)	0.9332	0.8351
MAE (train)	0.8002	0.7943
MAE (test)	0.8130	0.7807

Cuadro 1: Resultados Dataset 1

Métrica	RL	K-NN
R^2 (train)	0.3321	0.4157
R^2 (test)	0.3129	0.3630
MSE (train)	0.1092	0.0955
MSE (test)	0.1182	0.1096
MAE (train)	0.2486	0.2354
MAE (test)	0.2533	0.2456

Cuadro 2: Resultados Dataset 2

Métrica	RL	K-NN
R^2 (train)	0.7558	0.9002
R^2 (test)	0.7658	0.8123
MSE (train)	15232640	6224702
MSE (test)	14424950	11564542
MAE (train)	2776.11	1696.69
MAE (test)	2907.49	2403.50

Cuadro 3: Resultados Dataset 3

Dataset	Intervalo de confianza, $\hat{y}$ es el dato que predice el modelo
1	$(\hat{y} \pm 0{,}9641)$
2	$(\hat{y} \pm 0{,}3438)$
3	$(\hat{y} \pm 3798{,}0192)$

Cuadro 4: Intervalos de confianza

ligeramente a la regresión lineal. Para el segundo conjunto de datos, se aplicó una transformación logarítmica para mejorar la normalidad de los datos, y se observó que KNN continuó mostrando un mejor rendimiento en comparación con la regresión lineal. En el tercer conjunto de datos, con múltiples características, KNN destacó significativamente en la predicción del precio, superando a la regresión lineal.

Es importante señalar que la elección entre regresión lineal y KNN depende en gran medida de la naturaleza y la complejidad de los datos. Mientras que la regresión lineal asume una relación lineal entre variables, KNN se adapta mejor a patrones no lineales y puede ser más efectivo en conjuntos de datos complejos.

6. Conclusiones y Trabajo Futuro

En este informe, se exploraron y compararon dos técnicas de modelado, la regresión lineal simple y KNN, como un primer acercamiento a la estadística inferencial y el aprendizaje automático. Se concluye que, en los conjuntos de datos analizados, KNN mostró un rendimiento superior en términos de R^2, MSE y MAE en comparación con la regresión lineal.

Para trabajos futuros, se sugiere investigar y aplicar técnicas de preprocesamiento de datos más avanzadas, así como explorar otros algoritmos de aprendizaje automático. Además, se

podría considerar la optimización de hiperparámetros para mejorar aún más el rendimiento de los modelos. La exploración de conjuntos de datos más grandes y diversos también podría proporcionar una comprensión más completa de las capacidades de estos modelos en diferentes escenarios.

En resumen, este informe proporciona una introducción práctica a la aplicación de regresión lineal y KNN en la predicción de variables dependientes. Sin embargo, hay oportunidades para expandir y mejorar estos resultados mediante técnicas más avanzadas y la exploración de conjuntos de datos más desafiantes.

Referencias

[1] Diogo N Cosenza, Lauri Korhonen, Matti Maltamo, Petteri Packalen, Jacob L Strunk, Erik Næsset, Terje Gobakken, Paula Soares, and Margarida Tomé. Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock. *Forestry: An International Journal of Forest Research*, 94(2):311–323, 2021.

[2] Jorge Dagnino. Coeficiente de correlación lineal de pearson. *Chil Anest*, 43(1):150–153, 2014.

[3] Jorge Dagnino et al. Regresión lineal. *Revista chilena de anestesia*, 43(2):143–149, 2014.

[4] Elizabeth María Gandica de Roa. Potencia y robustez en pruebas de normalidad con simulación montecarlo. *Revista Scientific*, 5(18):108–119, 2020.

[5] Yue Shi. Application of improved linear regression algorithm in business behavior analysis. *Procedia Computer Science*, 228:1101–1109, 2023.

[6] Yunsheng Song, Jiye Liang, Jing Lu, and Xingwang Zhao. An efficient instance selection algorithm for k nearest neighbor regression. *Neurocomputing*, 251:26–34, 2017.

7. Enlace al Repositorio

Puedes encontrar el código fuente en GitHub:
`https://github.com/RodrigoCervantes-Data-AI-Eng/practice1_data-science.git`