PREDICCIÓN DEL PRECIO DE ALOJAMIENTOS EN LOS PAÍSES BAJOS

Miguel Ángel Sánchez Peñates

Daniel Góngora García

Jhon Alexander Longas

Tutor

Raúl Ramos Pollan, Professor of Computer Science

Introducción a la Inteligencia Artificial para las Ciencias e Ingeniería

Universidad de Antioquia

Facultad de Ingeniería

Ingeniería Industrial

Medellín, Antioquia, Colombia

2023

Contenido

Introducción

La inteligencia artificial (IA) sin duda alguna ha transformando la manera en que vemos y percibimos el mundo que nos rodea, y es que al tener grandes cantidades de datos en los diversos campos de las ciencias (economía, política, social, cultural, tecnológica, etc.) que son procesados y luego analizados para luego sacar sus respectivas conclusiones y tomar decisiones que contrarresten o en su de defecto, mejoren situaciones presentes y/o futuras relacionadas con los campos de las ciencias anteriormente mencionadas y esto se logran a través de una de las aplicaciones más importantes en el campo de la IA como lo es la predicción de datos.

La capacidad de predecir eventos futuros con un alto porcentaje de precisión es de gran importancia para la toma de decisiones en amplios sectores, desde la industrial y el comercio hasta la medicina y la meteorología. La importancia que tiene el predecir comportamientos que pueden llegar a suceder en un futuro por medio de las inteligencias artificiales radica en su capacidad para encontrar patrones y/o tendencias dentro de las grandes cantidades de datos y que para la mente humana muchas veces es difícil comprender o intuir, porque son en demasiadas ocasiones datos ocultos y tan complejos, que en todas las ocasiones nos quieren decir algo.

Es por eso que a través del aprendizaje automático (Machine Learning) y algoritmos sofisticados, los sistemas de la IA pueden explorar, estudiar y analizar grandes volúmenes de información y generar modelos que pronostican resultados futuros con una precisión cada vez mayor y confiable.

Esos modelos de predicción fundamentados en las inteligencias artificiales pueden tomar todos los datos históricos y variables relevantes para predecir eventos o comportamientos futuros. Por ejemplo, en el campo de la publicidad o marketing, las empresas pueden usar los datos históricos y a través de la IA para pronosticar el comportamiento de los clientes y cuáles son sus mayores necesidades y así poder ofrecer mejores servicios o productos y personalizar ofertas que se ajusten a las necesidades de un grupo de personas, aumentando así las ventas y, por ende, cumpliendo con las satisfacciones de los clientes.

Otro ejemplo claro sobre la importancia de las IA, se encuentra en el campo de la medicina. Y es que el poder predecir futuras enfermedades y cómo contrarrestarlas, es la tarea que muchos científicos y médicos han buscado y que a través de la generación de modelos de IA

han encontrado y eso se debe al análisis de datos de cada paciente, como la enfermedad que posee, los síntomas que tiene, los medicamentos que ha tomado y los resultados de pruebas para predecir el desarrollo de enfermedades y que los médicos puedan tomar decisiones más acertadas en cuanto a salvar vidas se trata.

En resumen, con el avance de la inteligencia artificial se ha mejorado significativamente nuestra capacidad para tomar mejores decisiones, en cuanto al examinar el comportamiento o las tendencias futuras de algunas situaciones o eventos que suceden en la sociedad. Al aprovechar el poder del aprendizaje automático y algunos algoritmos avanzados, la IA nos brinda la capacidad de analizar grandes volúmenes de datos, resumiéndola en información más explícita o detallada para las personas y observar los patrones ocultos, que permitan la generación de modelos predictivos más precisos. Esto tiene un gran impacto en la toma de decisiones estratégicas, la eficiencia operativa y el avance en materia de manufactura e industrial, lo que refleja la importancia de la inteligencia artificial en la predicción de datos.

Planteamiento del Problema

La crisis de la vivienda holandesa es uno de los mayores problemas a los que se enfrentan los residentes del país. La disponibilidad limitada de viviendas debido a múltiples factores, como el crecimiento de la población y la escasez de trabajadores de la construcción ha generado una disminución significativa en el acceso a viviendas adecuadas. Como resultado, los precios de alquiler han aumentado considerablemente, lo que despierta preocupación entre los desean alquilar una vivienda acerca de si están siendo explotados por los propietarios o las empresas inmobiliarias.

Para abordar esta cuestión y proporcionar una respuesta fundamentada, es necesario desarrollar un sistema de predicción de alquileres residenciales que utilice datos relevantes, como la ubicación, el tamaño de la vivienda, las instalaciones y otros factores determinantes. Este sistema de predicción permitirá estimar de manera precisa y objetiva el precio de alquiler de una casa en base a sus características, ayudando a las personas a tomar decisiones informadas en relación con sus opciones de vivienda.

El objetivo principal de este estudio es desarrollar un modelo de inteligencia artificial que sea capaz de predecir el precio de alquiler de una vivienda en los Países Bajos, que al utilizar datos históricos de alquiler y una variedad de variables, como la ubicación geográfica,

tamaño, servicios cercanos, infraestructura, entre otros, se pueda establecer un método preciso y confiable para estimar el precio justo del alquiler.

En resumen, el modelo de predicción ayudaría a eliminar o mitigar los efectos negativos de la crisis de vivienda en este país europeo al brindar una herramienta basada en IA que promueva la equidad y transparencia en el mercado de alquiler residencial y que pueda servir como base para la implementación de políticas y regulaciones más efectivas para abordar este problema y garantizar un acceso justo y asequible a la vivienda para los habitantes de Países Bajos.

Dataset o Base de Datos

El dataset seleccionado es de una competición de Kaggle llamada Precios de Alojamientos en Países Bajos, la cual podemos consultar en el siguiente enlace:

https://www.kaggle.com/competitions/fcg-2022-netherlands-accommodation-prices/overview.

Los datos de esta competición se han dividido en dos grupos:

- conjunto de entrenamiento (train.csv)
- conjunto de prueba (test.csv)

El conjunto de entrenamiento debe usarse para construir el modelo de aprendizaje automático. Para el conjunto de entrenamiento, proporcionamos el alquiler de cada alojamiento junto con otras 35 características.

El conjunto de pruebas debe usarse para ver cómo se desempeña el modelo en datos no vistos. Por lo tanto, no se proporciona el alquiler de cada alojamiento. El propósito del modelo es predecir estos valores.

Para ilustrar el formato de un archivo de envío, proporcionamos **sample_submission.csv**

La carpeta contiene los siguientes archivos

- **train.csv** - el conjunto de entrenamiento
- **test.csv** - el conjunto de prueba
- **sample_submission.csv**: un archivo de envío de muestra en el formato correcto

El archivo contiene las siguientes variables:

Variables	***Descripción***
Titulo	Nombre del alojamiento
Ciudad	Nombre de la ciudad
Código postal	Código postal
Latitud	Latitud en grados
Longitud	Longitud en grados
Área m^2	Tamaño en metros cuadrados
Visto por primera vez	Hora de registro del titular (AAAA-MM-DD HH-MM-SS) GMT
Visto por última vez	Última aparición del propietario (AAAA-MM-DD HH-MM-SS) GMT
isRoomActive	Disponibilidad actual
rawDisponibilidad	Periodo de tiempo de disponibilidad (DD-MM-AAAA)
Publicado hace	Hace cuánto tiempo se planteó la propiedad
Descripción no traducida	Descripción original
Descripción Traducido	Descripción traducida
Limpiar Detalle	Justificación del alquiler
Tipo de propiedad	Tipo de alojamiento
Amoblar	Presencia de muebles
Etiqueta de energía	Eficiencia de energética
Género	Sexo del propietario
Internet	Disponibilidad de internet
Compañeros de cuarto	Número de compañeros de cuarto
Ducha	Propiedad de la ducha
Baño	Propiedad del baño
Cocina	Propiedad de la cocina
Viviendo	Propiedad de la sala de estar
Mascotas	Mascotas permitidas

Fumar	Fumar está permitido
Edad	Edad permitida del inquilino
Coincidencia de genero	Sexo del inquilino deseado
Capacidad de coincidencia	# de personas que pueden vivir en el alojamiento
Coincidir idioma	Idioma deseado
Estado	Estado deseado
coverImageUrl	Url de la imagen de portada del alojamiento
Alquilar	Función objetivo

Tabla 1 Variables

Métrica de evaluación

La métrica de evaluación para el modelo será el Error Absoluto Promedio (MAE) el cual nos proporcionará el promedio de la diferencia absoluta entre la predicción del modelo y el valor objetivo.

Esta métrica se calcula de la siguiente manera:

$$MAE = \frac{\left(\sum_{i=1}^{n} \left|y_i - \underline{y_i}\right|\right)}{n}$$

Donde:

$y_i = son\ la\ observaciones\ actuales\ de\ las\ series\ de\ tiempo.$

$\underline{y_i} = es\ la\ serie\ de\ tiempo\ estimada\ o\ pronosticada.$

$n = es\ el\ número\ de\ puntos\ de\ datos\ no\ fal\ tan\ tan\ t\ es$

Es importante resaltar que el MAE lo siguiente:

- Evaluación de precisión: El MAE permite evaluar cuán cerca están las predicciones del modelo del valor real del alquiler. Un MAE más bajo indica que las predicciones son más precisas y se acercan más a los valores reales.
- Comparación de modelos: Al calcular el MAE para diferentes modelos de predicción, se puede comparar su desempeño y determinar cual ofrece predicciones más precisas. Un modelo con un MAE más bajo sería considerado más confiable y efectivo en la estimación de los precios de alquiler.

- Mejora el modelo: El MAE también puede ser usado para optimizar y mejorar los modelos de predicción. Al identificar las características o variables que contribuyen significativamente a un mayor error absoluto, se pueden realizar ajustes y mejoras en el modelo para reducir el error y aumentar la precisión de las predicciones.
- Transparencia y confianza: Al comunicar el error absoluto promedio al público y los usuarios del sistema de predicción, se fomenta la transparencia y se establece una medida objetiva para evaluar la confiabilidad del modelo. Esto genera mayor confianza en las estimaciones de alquiler y facilita la toma de decisiones informadas por parte de los inquilinos y propietarios.

Desempeño

El desempeño esperado de un modelo para pronosticar es que el precio de los alojamientos respectos a los datos de entrenamiento es que exista una tendencia a la baja, lo que nos representaría que el modelo es adecuado para determinar si los precios de los alquileres están ajustados a las características de cada alojamiento, ya que dichos precios pueden varias según varios factores, como la disponibilidad y la calidad de los datos. Por tal razón, se espera que el modelo sea capaz de generar predicciones precisas y cercanas al valor real del precio de los alojamientos. Esto implicaría minimizar el error absoluto promedio y otros indicadores de error, lo que significa que las predicciones se ajustan de manera confiable a los precios reales observados.

Por ejemplo, si el modelo predice un precio de un alquiler es 1200€ cuando el precio real es de 1000€, el error es del 20%. Por otro lado, si el precio real es de 400€ y el modelo predice 200€, el error es del 50%. Utilizar el porcentaje de error como métrica para evaluar el desempeño del modelo es una aproximación valida. Esto implica evaluar el impacto del modelo en términos de mejorar la toma de decisiones de precios y aumentar la demanda de los alojamientos.

Análisis Exploratorio de los datos

En fase de análisis y comprensión inicial de los datos se realiza el EDA (Análisis exploratorio de los datos) se extrae información importante antes de realizar un modelo más avanzado de los datos, permitiendo examinar la distribución de las variables. Se pueden identificar

características como la simetría, la presencia de valores atípicos y la necesidad de transformar las variables para cumplir con los supuestos de algunos modelos. Además, se realiza un análisis de correlación entre las variables que revelan las relaciones lineales entre las variables, permitiendo comprender las interacciones entre las características más relevantes para el modelado.

Además, al analizar los datos, es necesario realizarles un tratamiento a los datos faltantes, lo que permite determinar la cantidad y ubicación de los valores faltantes y que, a su vez, permite decidir cómo se maneja esos valores nulos, ya sea eliminándolos o utilizando técnicas más avanzadas de manejo de datos faltantes. Asimismo, se realiza un análisis univariable, lo que nos permite examinar de forma individual cada variable e identificar patrones o tendencias. Por ejemplo, se puede observar la distribución de los precios de alojamientos y obtener información sobre su dispersión y valores centrales.

Para llevar a cabo el EDA se siguieron las siguientes etapas:

Obtención de los datos: Como se mencionó con anterioridad, los datos fueron recolectados de una competición de Kaggle y lo que se busca con dicha base de datos es predecir el valor de los precios de alojamiento en Países Bajos.

Cargamos la base de datos

Los datos se cargan en un entorno de análisis como lo es Google Colab, como un DataFrame en Python, utilizando, por ejemplo, la biblioteca de Pandas. El resultado es el siguiente:

id	title	city	postalCode	latitude	longitude	areaSqm	firstSeenAt	lastSeenAt	isRoomActive	rawAvailability	...	living
0	West-Varkenoordseweg	Rotterdam	3074HN	51.896601	4.514993	14	2019-07-14 11:25:46.511000+00:00	2019-07-26 22:18:23.142000+00:00	True	26-06-'19 - Indefinite period	...	None
3	Ruiterakker	Assen	9407BG	53.013494	6.561012	16	2019-07-14 11:25:46.988000+00:00	2019-07-18 22:00:31.174000+00:00	False	16-06-'19 - Indefinite period	...	None
8	Brusselseweg	Maastricht	6217GX	50.860841	5.671673	16	2019-07-14 11:25:47.814000+00:00	2019-08-10 00:14:27.130000+00:00	True	15-07-'19 - Indefinite period	...	None
10	Donkerslootstraat	Rotterdam	3074WL	51.893195	4.516478	25	2019-07-14 11:25:48.140000+00:00	2019-07-16 06:05:32.183000+00:00	False	01-08-'19 - Indefinite period	...	None
12	Vorselenburgstraat	Alphen aan den Rijn	2405XJ	52.122335	4.661434	10	2019-07-14 11:25:48.465000+00:00	2019-08-01 00:02:40.516000+00:00	True	08-07-'19 - Indefinite period	...	None

5 rows × 33 columns

Figura 1 Datos

Número de observaciones y valores ausentes

Junto con el estudio del tipo de variables, es básico conocer el número de observaciones disponibles y si todas ellas están completas. Los valores ausentes son muy importantes a la hora de crear modelos, la mayoría de los algoritmos no aceptan observaciones incompletas o bien se ven muy influenciados por ellas. Aunque la imputación de valores ausentes es parte del preprocesado y, por lo tanto, debe de aprenderse únicamente con los datos de entrenamiento, su identificación se tiene que realizar antes de separar los datos para asegurar que se establecen todas las estrategias de imputación necesarias.

- Dimensión de la base de datos: (27915, 33).
- Número de datos ausentes por variables: Variables como gender, roommates, rentDetail, descriptionTranslated son las que más variables faltantes tienen.

```
title                          0
coverImageUrl                  0
propertyType                   0
rawAvailability                0
lastSeenAt                     0
firstSeenAt                    0
rent                           0
longitude                      0
latitude                       0
postalCode                     0
city                           0
areaSqm                        0
postedAgo                      6
matchStatus                   63
matchLanguages                63
matchCapacity                 63
matchGender                   63
matchAge                      63
smokingInside                 63
pets                          63
living                        63
kitchen                       63
energyLabel                   63
shower                        63
internet                      63
isRoomActive                  63
toilet                        63
descriptionNonTranslated     111
furnish                      214
gender                       536
roommates                    536
rentDetail                  7896
descriptionTranslated      10140
dtype: int64
```

Análisis de la variable respuesta

Cuando se crea un modelo, es muy importante estudiar la distribución de la variable respuesta, ya que, a fin de cuentas, es lo que interesa predecir. La variable alquiler (rent) tiene una distribución asimétrica con una cola positiva debido a que, unos pocos alojamientos, tienen un precio muy superior a la media. Este tipo de distribución suele visualizarse mejor si se calcular la asimetría de los datos, que en este caso es 2.36, esto sugiere entonces que hay valores atípicos en el lado derecho de la distribución, comprobando entonces que hay presencia de valores de alquileres más altos, por lo cual podemos concluir que presenta una distribución no normal. ***Ver figura 3***

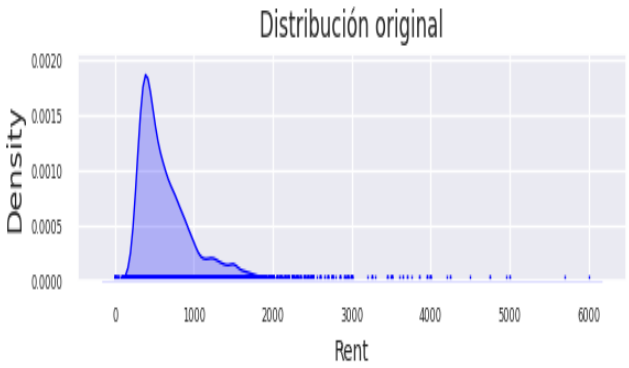


Figura 3 Variable Objetivo

Para normalizar la variable respuesta se realizó una transformación de dicha variable ya que anteriormente mostraba un gran desfase y como muestra la ***figura 4*** vemos que la curva de la distribución de asemeja más a la curva de la normal, lo que quiere decir que la distribución se vuelve más simétrica y cercana a cero indicando que los datos están distribuidos de manera más equilibrada alrededor de la media.

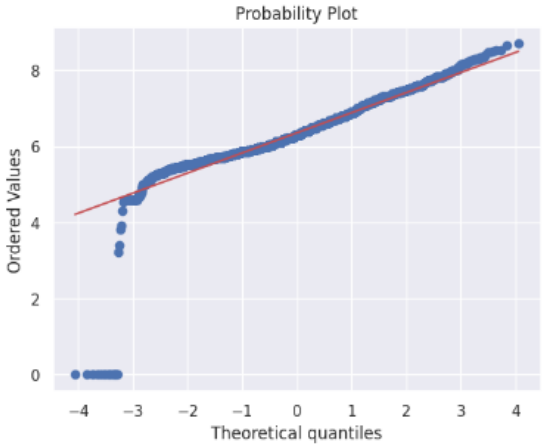


Figura 4 Distribución normal de la variable objetivo

Descripción de las variables numéricas

Se realizó los gráficos de la distribución de cada una de las variables numéricas y observamos que los datos tanto de la variable Latitud como Longitud presentan una distribución normal, ya que casi todos los datos están alrededor de la media, es decir, que mucha de la demanda

de alquileres se concentra en ciudades como Ámsterdam, Utrecht o incluso Deventer. ***Ver figura 5***

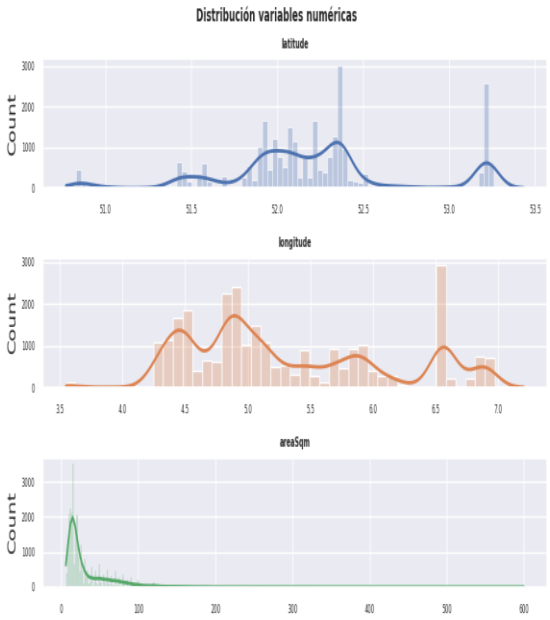


Figura 5 Distribución Variables numéricas

Como el objetivo del estudio es predecir el precio del alquiler de los alojamientos, el análisis de cada variable se hace también en relación con la variable respuesta precio. Al analizar los datos podemos observar en la ***figura 6***, que los datos se concentran más en las ciudades con latitud 52º y longitud 5º y donde los alojamientos más buscados son aquellos con un área menor a los 100m^2

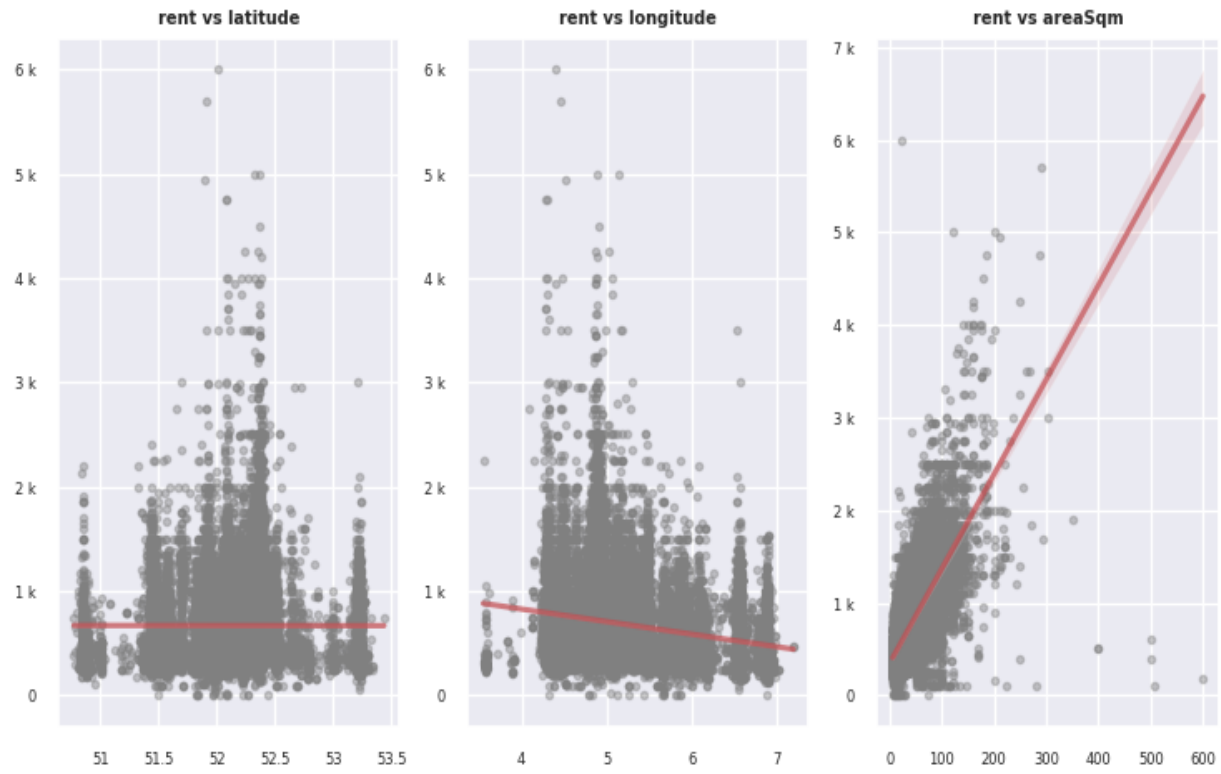


Figura 6 Relación variable respuesta vs variables numéricas

Correlación de las variables numéricas

Al observar la figura 7 podemos sacar las siguientes conclusiones:

- Existe una correlación positiva fuerte entre estas entre la variable respuesta y la variable área, con un valor de correlación de 0.729018. Esto indica que a medida que el tamaño del área (areaSqm) aumenta, es probable que el precio del alquiler (rent) también aumente. Esta relación tiene sentido, ya que es razonable esperar que los alquileres sean más altos para áreas más grandes.
- Existe una correlación moderada entre estas dos variables latitud y longitud, con un valor de correlación de 0.405798. Esto sugiere que existe una relación entre la latitud y la longitud de la ubicación de las propiedades. Sin embargo, es importante tener en cuenta que la correlación entre estas variables no indica una relación causal directa, sino más bien una tendencia general.

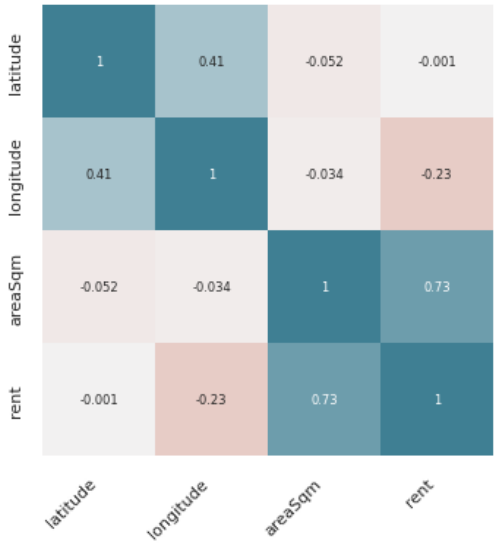


Figura 7 Correlación entre las variables

- Hay una correlación débil negativa entre la variable respuesta y la variable longitud, con un valor de correlación de -0.230255. Esto implica que a medida que aumenta la longitud geográfica, es posible que el precio del alquiler disminuya ligeramente. Sin embargo, la correlación es débil, lo que sugiere que la relación entre estas variables puede no ser muy significativa.
- Al igual que en el punto anterior, se observa una correlación débil negativa entre la variable respuesta y la variable latitud, con un valor de correlación de -0.230255. Esto indica que a medida que aumenta la latitud geográfica, el precio del alquiler podría disminuir ligeramente.

Datos faltantes

Es importante analizar antes de entrenar un código de machine learning, cuáles son las variables que poseen los mayores datos faltantes ya que estos pueden afectar la calidad de los datos y la validez de los resultados del análisis. Si hay demasiados datos faltantes en una variable, puede ser difícil o imposible obtener conclusiones precisas o representativas sobre esa variable. En la figura 8 vemos como alrededor del 36,3% de los datos en la variable “descriptionTranslated” están sin datos y que alrededor del 28.3% de la variable “rentDetail” está también con datos faltantes, por lo cual se tienen que tomar medidas que haga estas variables no afecten la precisión y la calidad de los resultados esperados.

	Total	Percent
descriptionTranslated	10140	0.363246
rentDetail	7896	0.282859
roommates	536	0.019201
gender	536	0.019201
furnish	214	0.007666
descriptionNonTranslated	111	0.003976
kitchen	63	0.002257
shower	63	0.002257
internet	63	0.002257
living	63	0.002257
pets	63	0.002257
energyLabel	63	0.002257

Figura 8 Datos Faltantes por variable

Preprocesamiento de los datos

Según lo observado en el análisis exploratorio, se realizó un preprocesamiento bastante básico que incluye:

- **Eliminación de datos nulos:** Del análisis exploratorio de las variables, se encontró que existen variables que contienen gran cantidad de datos faltantes. En este caso se optará por eliminar las columnas en las que los datos faltantes representen el 80% o más de la totalidad de los datos. Siguiendo este criterio se eliminan las variables "rentDetail", "descriptionTranslated ". Tras esto las dimensiones del dataset se redujeron a *27915 x 31*.
- **Transformación de las variables categóricas:** Las categorías que se pueden interpretar como números, fueron transformadas a variables numéricas, por ejemplo la variable *roommates*, a la cual también se le aplicó un método de rellenado de datos nulos, cambiandolas con la media global para todo el dataset.

Métodos Supervisados

Selección del modelo

- Para evaluar los resultados de los modelos supervisados, se utilizó el error absoluto medio (MAE) como métrica de evaluación. El MAE representa la diferencia promedio entre los valores reales y los valores predichos por el modelo. Un MAE más bajo indica un mejor rendimiento del modelo, ya que implica una menor diferencia entre las predicciones y los valores reales.

  Comparando los resultados de los tres modelos:

1. Modelo de Regresión Lineal:

- MAE: 0.1526
- Ejemplos de predicciones:
  - Para la observación con id 39236, el valor real es 5.98 y el valor predicho es 6.01.
  - Para la observación con id 6949, el valor real es 6.62 y el valor predicho es 6.56.
  - Para la observación con id 38286, el valor real es 7.24 y el valor predicho es 7.15.

2. Modelo SVM:

- MAE: 0.1804
- Ejemplos de predicciones:
  - Para la observación con id 12551, el valor real es 8.15 y el valor predicho es 7.72.
  - Para la observación con id 8881, el valor real es 6.21 y el valor predicho es 6.05.
  - Para la observación con id 13694, el valor real es 7.21 y el valor predicho es 7.06.

3. Modelo de Árbol de Decisión:

- MAE: 0.1684
- Ejemplos de predicciones:
    - Para la observación con id 5322, el valor real es 6.01 y el valor predicho es 5.62.
    - Para la observación con id 38405, el valor real es 6.02 y el valor predicho es 6.02.
    - Para la observación con id 15761, el valor real es 6.40 y el valor predicho es 6.40.

El modelo de regresión lineal muestra el menor MAE en comparación con los otros dos modelos. Esto indica que el modelo de regresión lineal tiene un mejor rendimiento en la predicción de los valores de la variable objetivo (rent) en comparación con los modelos SVM y de árbol de decisión. El MAE más bajo indica que las predicciones del modelo de regresión lineal están más cerca de los valores reales en promedio.

Por lo tanto, se podría concluir que el modelo de regresión lineal es el mejor modelo entre los tres evaluados en términos de su capacidad para predecir los valores de la variable objetivo. Sin embargo, es importante tener en cuenta que esta conclusión se basa únicamente en la métrica del MAE y que otros factores, como la complejidad del modelo y la interpretación de los coeficientes, también deben considerarse al seleccionar el mejor modelo para un problema específico.

Resultados y Métricas

Los resultados y las métricas obtenidas de los modelos utilizados para pronosticar el precio de alojamientos en países bajos pueden proporcionar información valiosa sobre el desempeño de los modelos y su capacidad para realizar predicciones precisas. Al analizar los resultados y las métricas, se pueden extraer las siguientes conclusiones:

1. Resultados de las predicciones: Los resultados muestran las comparaciones entre los valores reales del precio de alojamiento y los valores predichos por cada modelo. Es importante observar si las predicciones están cercanas a los valores reales, ya que esto indica la precisión del modelo en la estimación del precio. Si los valores predichos están en línea

con los valores reales, se puede considerar que el modelo está realizando buenas predicciones.

2. Métricas de evaluación: En este caso, se utilizó el MAE (Error Absoluto Medio) como métrica de evaluación para medir la diferencia promedio entre los valores reales y los valores predichos por cada modelo. Un MAE bajo indica que las predicciones del modelo son cercanas en promedio a los valores reales. Por lo tanto, cuanto más bajo sea el valor del MAE, mejor será el desempeño del modelo.

En base a los resultados y las métricas proporcionadas, se puede decir lo siguiente:

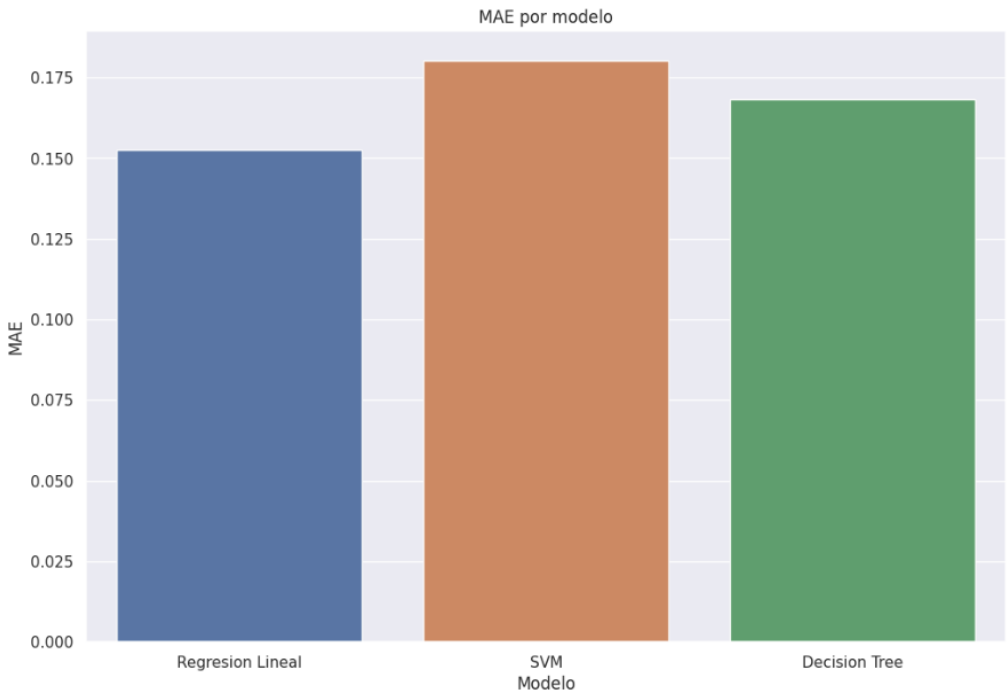
- El modelo de regresión lineal muestra un MAE de 0.1526, lo que indica que las predicciones del modelo están en promedio cercanas al precio real de los alojamientos. Esto sugiere que el modelo de regresión lineal tiene un buen desempeño y es capaz de realizar estimaciones precisas del precio de alquiler.

- El modelo SVM tiene un MAE de 0.1804, lo que indica una ligera diferencia promedio entre las predicciones y los valores reales. Aunque el MAE es un poco más alto en comparación con el modelo de regresión lineal, aún se considera aceptable y sugiere que el modelo SVM también puede realizar predicciones razonablemente precisas.

- El modelo de árbol de decisión muestra un MAE de 0.1684, que se encuentra en un rango similar al MAE del modelo SVM. Esto indica que el modelo de árbol de decisión tiene una capacidad moderada para realizar estimaciones del precio de alojamiento, aunque puede haber una ligera diferencia promedio entre las predicciones y los valores reales.

En resumen, los resultados y las métricas indican que los modelos utilizados tienen un desempeño razonable en la predicción del precio de alojamientos en países bajos. Sin embargo, es importante considerar que cada modelo tiene sus propias fortalezas y limitaciones, y es recomendable evaluar diferentes métricas y realizar una comparación exhaustiva antes de tomar una decisión sobre el mejor modelo a utilizar.

La métrica utilizada para ‘calificar’ los modelos desplegados fue el MAE (Error Absoluto Medio) y considerando los 3 modelos se obtuvieron los siguientes resultados:



Con base en esto, se sugiere la implementación de un modelo de regresión lineal para la predicción del precio de los alojamientos en este caso de estudio.

Retos y despliegue del modelo

El modelo de regresión lineal puede enfrentar varios desafíos y consideraciones al predecir precios de alquileres. Algunos de los desafíos y despliegues comunes son:

- Datos no lineales: El modelo de regresión lineal asume una relación lineal entre las variables independientes y la variable objetivo. Sin embargo, en el caso de los precios de alquileres, es probable que existan relaciones no lineales, como efectos de umbral o interacciones no lineales. En tales casos, el modelo de regresión lineal puede no capturar adecuadamente estas relaciones y producir predicciones inexactas.
- Variables irrelevantes: El modelo de regresión lineal puede verse afectado por la inclusión de variables irrelevantes o altamente correlacionadas en el conjunto de características. Esto puede conducir a una estimación sesgada de los coeficientes y

afectar la precisión de las predicciones. Es importante realizar un análisis cuidadoso de las variables y seleccionar aquellas que tengan una influencia significativa en el precio de alquiler.

- Multicolinealidad: La multicolinealidad ocurre cuando hay alta correlación entre las variables independientes. Esto puede dificultar la interpretación de los coeficientes y aumentar la varianza de las estimaciones. En presencia de multicolinealidad, se pueden tomar medidas como eliminar variables altamente correlacionadas o utilizar técnicas de regularización para mitigar este problema.
- Extrapolación: El modelo de regresión lineal se basa en la suposición de que la relación entre las variables independientes y la variable objetivo se mantiene constante dentro del rango de los datos de entrenamiento. Sin embargo, al realizar predicciones fuera de este rango, se está extrapolando y existe el riesgo de que las predicciones sean menos precisas. Es importante tener en cuenta las limitaciones de extrapolación al interpretar las predicciones del modelo.
- Supuestos de la regresión lineal: El modelo de regresión lineal se basa en ciertos supuestos, como la linealidad, la independencia de errores, la normalidad de los errores y la homocedasticidad. Es importante verificar si estos supuestos se cumplen antes de confiar en las predicciones del modelo. En caso contrario, pueden requerir técnicas alternativas, como la regresión lineal robusta o modelos no lineales.
- Actualización de datos: Los precios de alquileres pueden estar sujetos a cambios frecuentes debido a factores económicos, estacionales o de mercado. Por lo tanto, es importante tener un mecanismo para actualizar regularmente los datos utilizados para entrenar el modelo y garantizar que esté capturando las tendencias actuales.

En resumen, el modelo de regresión lineal puede enfrentar desafíos en la predicción de precios de alquileres debido a la naturaleza no lineal de los datos, variables irrelevantes, multicolinealidad y supuestos de la regresión lineal. Es necesario abordar estos desafíos y considerar técnicas más avanzadas si es necesario para mejorar la precisión y la interpretación de las predicciones.

Conclusiones

Algunas conclusiones que se pueden extraer de los modelos utilizados para pronosticar el precio de alojamientos o alquiler en países bajos son las siguientes:

1. Modelo de regresión lineal: El modelo de regresión lineal muestra un buen desempeño en la predicción del precio de alojamientos en países bajos, con un MAE (Error Absoluto Medio) de 0.1526. Esto indica que las predicciones del modelo están cercanas en promedio al precio real de los alojamientos. Sin embargo, se debe tener en cuenta que el modelo de regresión lineal asume una relación lineal entre las variables independientes y el precio de alquiler, lo que puede limitar su capacidad para capturar relaciones no lineales más complejas.

2. Modelo SVM (Máquinas de Vectores de Soporte): El modelo SVM también muestra un desempeño aceptable en la predicción del precio de alojamientos en países bajos, con un MAE de 0.1804. Las SVM son capaces de capturar relaciones no lineales mediante el uso de funciones de kernel, lo que puede permitir una mejor adaptación a los datos. Sin embargo, es importante tener en cuenta que el modelo SVM puede ser más computacionalmente intensivo y requiere una optimización adecuada de los hiperparámetros.

3. Modelo de árbol de decisión: El modelo de árbol de decisión presenta un MAE de 0.1684, lo que indica una capacidad moderada para predecir el precio de alojamientos en países bajos. Los árboles de decisión son modelos no lineales que pueden capturar relaciones complejas entre las variables independientes y la variable objetivo. Sin embargo, los árboles de decisión también pueden ser propensos al sobreajuste y pueden generar resultados menos estables en comparación con otros modelos.

En general, se puede concluir que los modelos utilizados muestran una capacidad razonable para pronosticar el precio de alojamientos en países bajos. Sin embargo, es importante considerar las limitaciones de cada modelo y tener en cuenta que otros factores, como la ubicación geográfica específica, las características del alojamiento y las condiciones del mercado, pueden influir en el precio y no están considerados en estos modelos. Por lo tanto, se recomienda utilizar estos modelos como herramientas complementarias en la estimación

del precio de alojamientos y considerar otros factores relevantes para obtener pronósticos más precisos.

Bibliografía

Netherlands Accommodation Prices (FCG) | Kaggle. (s. f.). https://www.kaggle.com/competitions/fcg-2022-netherlands-accommodation-prices/overview/evaluation