**DOI:** https://doi.org/10.23857/dc.v10i3.4022

Ciencias Técnicas y Aplicadas
Artículo de Investigación

# *Implementación del Método de Análisis de Componentes Principales (PCA) para la reducción de la dimensionalidad en los datos inmobiliarios de la ciudad de Riobamba*

# *Implementation of the Principal Component Analysis (PCA) method for dimensionality reduction in real estate data of the city of Riobamba*

# *Implementação do Método de Análise de Componentes Principais (PCA) para a redução de dimensionalidade em dados imobiliários da cidade de Riobamba*

Carlos Antonio Aguirre Calderón [I]
antonio.aguirre@unach.edu.ec
https://orcid.org/0009-0006-7396-7759

Elba María Bodero Poveda [II]
ebodero@unach.edu.ec
https://orcid.org/0000-0003-3807-5203

**Correspondencia:** antonio.aguirre@unach.edu.ec

*__Recibido:__ 11 de agosto de 2024 *__Aceptado:__ 02 de septiembre de 2024 __* Publicado:__ 17 de septiembre de 2024

I. Ingeniero Civil. Maestrante en la Universidad Nacional de Chimborazo. Riobamba-Ecuador.
II. Doctora en Ciencias Informáticas. Docente e Investigadora del Grupo de Investigación en Telecomunicaciones, Informática, Industria y Construcción en la Universidad Nacional de Chimborazo. Riobamba-Ecuador.

## Resumen

El acelerado crecimiento inmobiliario en la ciudad de Riobamba ha generado un gran volumen de datos con desafíos de dimensionalidad y complejidad. En busca de una posible mejora del rendimiento en la implementación de modelos computacionales, esta investigación tiene como objetivo aplicar el método Análisis de Componentes Principales (PCA) para reducir la dimensionalidad en los datos inmobiliarios de la ciudad de Riobamba. La población de estudio comprende la información inmobiliaria de 2338 propiedades (bienes raíces) de la ciudad. El muestreo es de tipo censal, incluyen todos los registros del periodo 2018-2022. La investigación tiene un alcance explicativo, con un diseño cuasi – experimental. La investigación tiene un alcance explicativo y un diseño cuasi – experimental. Los resultados obtenidos muestran que, al aplicar PCA, la reducción de dimensionalidad fue satisfactoria para las variables numéricas. Al analizar los datos transformados, se observó que los componentes principales retenían un porcentaje significativo de la variabilidad total de los datos originales. Esto indica que las variables originales pueden ser representadas adecuadamente con un número reducido de componentes principales, lo que simplifica el modelo sin perder información relevante.

**Palabras Claves:** Análisis de Componentes Principales PCA; Datos inmobiliarios; Reducción de la Dimensionalidad.

## Abstract

The accelerated real estate growth in the city of Riobamba has generated a large volume of data with dimensionality and complexity challenges. In search of a possible performance improvement in the implementation of computational models, this research aims to apply the Principal Component Analysis (PCA) method to reduce dimensionality in the real estate data of the city of Riobamba. The study population comprises the real estate information of 2338 properties (real estate) in the city. The sampling is census-type, including all records from the period 2018-2022. The research has an explanatory scope, with a quasi-experimental design. The research has an explanatory scope and a quasi-experimental design. The results obtained show that, when applying PCA, the dimensionality reduction was satisfactory for the numerical variables. When analyzing the transformed data, it was observed that the principal components retained a significant percentage of the total variability of the

original data. This indicates that the original variables can be adequately represented with a reduced number of principal components, which simplifies the model without losing relevant information.
**Keywords:** Principal Component Analysis PCA; Real estate data; Dimensionality reduction.

## Resumo

O acelerado crescimento imobiliário na cidade de Riobamba gerou um grande volume de dados com desafios de dimensionalidade e complexidade. Na procura de uma possível melhoria de desempenho na implementação de modelos computacionais, esta investigação tem como objetivo aplicar o método Análise de Componentes Principais (PCA) para a redução de dimensionalidade em dados imobiliários da cidade de Riobamba. A população de estudo compreende a informação imobiliária de 2.338 imóveis (imóveis) do concelho. A amostragem é do tipo censitário, inclui todos os registos do período 2018-2022. A investigação tem um âmbito explicativo, com um desenho quase experimental. A investigação tem um âmbito explicativo e um desenho quase-experimental. Os resultados obtidos mostram que, ao aplicar o PCA, a redução da dimensionalidade foi satisfatória para as variáveis numéricas. Ao analisar os dados transformados, observou-se que os componentes principais retiveram uma percentagem significativa da variabilidade total dos dados originais. Isto indica que as variáveis originais podem ser representadas adequadamente com um número reduzido de componentes principais, o que simplifica o modelo sem perder informação relevante.
**Palavras-chave:** Análise de Componentes Principais PCA; Dados imobiliários; Redução de dimensionalidade.

## Introducción

En la era de la información, la cantidad de datos generados y recopilados en diversos campos ha aumentado exponencialmente, presentando tanto oportunidades como desafíos importantes para la investigación y la industria. El análisis de grandes volúmenes de datos se ha convertido en una herramienta fundamental para la toma de decisiones en múltiples áreas, incluido el sector inmobiliario. La gestión eficiente de los datos inmobiliarios puede ofrecer información valiosa a compradores, vendedores y autoridades municipales, favoreciendo el desarrollo urbano y la planificación estratégica.

El campo de la tecnología de la información ha experimentado un rápido crecimiento (Sarker, 2021). Con el objetivo de impulsar la mejora continua en empresas y organizaciones a nivel mundial

mediante el análisis de estos datos, los algoritmos de Aprendizaje Automático o Machine Learning (ML) se destacan como herramientas útiles para identificar patrones y tendencias. Sin embargo, es importante tener en cuenta que la aplicación de algoritmos de ML en el análisis de grandes conjuntos de datos puede enfrentar dificultades debido a la voluminosa cantidad de registros y columnas, lo que conlleva a un alto requerimiento de procesamiento computacional (Bodero et al., 2020a).

Pernice (2024) destaca que el Análisis de Componentes Principales (PCA, por sus siglas en inglés) ha tenido un impacto significativo en la investigación empírica en campos como la economía, las finanzas y otras ciencias sociales. Actualmente, el desarrollo de modelos para analizar datos de panel de alta dimensionalidad, que se basan directa o indirectamente en el PCA, constituye un área de investigación activa. El PCA desempeña un papel crucial al simplificar los datos y eliminar el ruido (Cheng et al., 2021). Este método es especialmente eficaz para medir tanto la tendencia como la similitud numérica entre secuencias de datos de alta y baja dimensión (Wang et al., 2024a).

Además, el PCA es ampliamente reconocido como el método más adoptado y aplicado en el análisis multivariado de datos. Como técnica estadística flexible, permite reducir una matriz de datos organizada por casos y variables a sus componentes principales, también conocidos como elementos esenciales (Kurita, 2019). Su prominencia se debe a su eficacia para reducir la complejidad de conjuntos de datos con múltiples variables, al identificar las principales direcciones de variabilidad y facilitar una comprensión más clara de la estructura subyacente de los datos (Gonzalez et al., 2021).

El proyecto "Minería de datos para la predicción de la plusvalía inmobiliaria en el cantón Riobamba", ejecutado en la Universidad Nacional de Chimborazo hasta el año 2022, permitió recopilar 2338 registros de propiedades inmobiliarias. Esta cantidad de datos presenta desafíos en términos de dimensionalidad y complejidad. En este contexto, surgió la necesidad de gestionar eficientemente estos datos para extraer información relevante y significativa (Bodero et al., 2022b). Así, el PCA se presentó como una herramienta potencial para reducir la dimensionalidad de los datos inmobiliarios, permitiendo una representación más compacta y comprensible de la información. Según Mostofi et al. (2022), el PCA ofrece beneficios clave para mejorar la precisión en la predicción de precios en el sector inmobiliario.

A pesar de la existencia de varios estudios sobre el PCA, la mayor parte de la literatura se enfoca en áreas distintas y no aborda de manera específica las aplicaciones y desafíos del PCA en el sector inmobiliario. Por lo tanto, es fundamental plantear la siguiente pregunta de investigación: ¿Cómo permite la implementación del método PCA reducir la dimensionalidad en los datos inmobiliarios de

la ciudad de Riobamba? Para responder a esta pregunta, es necesario realizar un análisis teórico y matemático del funcionamiento del PCA, evaluar la calidad de los datos inmobiliarios actualmente recopilados, aplicar el PCA mediante un algoritmo en un lenguaje de programación orientado al análisis de datos, y finalmente, validar su efectividad centrándose en el rendimiento.

En este contexto, se buscó comprender y analizar cómo la aplicación específica del método PCA impacta la reducción de dimensionalidad en conjuntos de datos relacionados con el sector inmobiliario en la ciudad de Riobamba. El propósito central de esta investigación fue obtener conocimientos detallados sobre las técnicas efectivas, los alcances, los desafíos y las aplicaciones prácticas del PCA en este contexto particular. La resolución de esta interrogante contribuirá a mejorar la interpretación y el análisis de datos inmobiliarios, facilitando una toma de decisiones más informada y estableciendo las bases para futuros avances y desarrollos en este campo.

## Metodología

El estudio tiene un enfoque cuantitativo y, según su alcance, es de tipo explicativo, debido a que busca establecer la efectividad de la aplicación del método PCA en la reducción de dimensionalidad de los datos inmobiliarios de la ciudad de Riobamba. Su diseño es cuasi-experimental, debido a la manipulación mínima de las variables. Además, es longitudinal, dado que se tomaron datos a lo largo de varios años. Se aplicará el lenguaje de programación Python tanto para realizar el análisis de calidad de datos, la aplicación de PCA y la validación de efectividad del método de reducción de dimensionalidad. Para este último objetivo se hará uso de:

- **Proporción de varianza explicada:** Se refiere a la cantidad de varianza en los datos originales que es explicada por las nuevas variables (componentes) generadas por el PCA (Jiménez et al., 2002). Cuanto mayor sea esta proporción, más efectivo será el método para resumir la información contenida en los datos originales.
- **Número óptimo de componentes:** El PCA genera un conjunto de componentes ordenados según la cantidad de varianza que explican. Determinar el número óptimo de componentes implica encontrar un equilibrio entre la cantidad de información retenida y la reducción de la dimensionalidad (Dray, 2008).
- **Rendimiento:** Esta métrica evalúa cómo se comportan los datos después de reducir su dimensionalidad utilizando PCA, mediante la aplicación de técnicas de ML (Cruz et al., 2022).

La población de estudio comprende la información inmobiliaria de 2338 propiedades (bienes raíces) de la ciudad de Riobamba. El muestreo es de tipo censal porque se incluyen en el análisis todos los registros.

La información inmobiliaria de la ciudad de Riobamba se encuentra en una base de datos que fue recolectada en el proyecto de investigación UNACH "Minería de datos para la predicción de la plusvalía inmobiliaria en el cantón Riobamba", los años que se trabajará en esta investigación corresponden al periodo desde 2018 al 2022. El procedimiento fue el siguiente:

Objetivo 1: Analizar la calidad de los datos inmobiliarios, a través de un análisis exploratorio descriptivo, identificación de valores atípicos, verificación de la integridad y la consistencia, con la finalidad de reducir el riesgo asociado a la utilización de datos erróneos.

1. Importación del conjunto de datos
2. Eliminación de filas y columnas duplicadas
3. Identificación y tratamiento de valores nulos
4. Identificación y eliminación de valores atípicos
5. Normalización y estandarización de los datos
6. Verificar la integridad y consistencia

Objetivo 2: Aplicar el método PCA, mediante un algoritmo informático, para reducir la dimensionalidad en los datos inmobiliarios de la ciudad de Riobamba.

1. Preparación de los datos para PCA
2. Separar el conjunto de datos en uno de variables numéricas y otro de categóricas
3. Codificar las variables categóricas
4. Estandarizar las variables para tener de media 0 y varianza 1
5. Aplicar PCA en variables numéricas
6. Combinar variables numéricas y categóricas
7. Aplicar PCA a datos combinados

Objetivo 3: Validar la efectividad del método PCA, a través de la proporción de varianza explicada, número óptimo de componentes y rendimiento, para establecer si la aplicación de este método permite una reducción óptima de la dimensionalidad de los datos inmobiliarios de la ciudad de Riobamba.

1. Análisis de la varianza explicada y varianza explicada acumulada
2. Identificar el número óptimo de componentes principales
3. Entrenar un modelo de regresión lineal

4. Evaluar el rendimiento del modelo
5. Visualizar los resultados obtenidos

## Desarrollo

### Análisis de la Calidad de los datos

Evaluar la calidad de los datos inmobiliarios de la ciudad de Riobamba es un proceso fundamental para asegurar la validez y precisión de la información utilizada en el análisis de PCA, con el objetivo de minimizar los riesgos asociados al uso de datos incorrectos o incompletos. Este proceso sistemático incluyó un análisis exploratorio descriptivo, la identificación de valores atípicos y la verificación de la integridad y consistencia de los datos. La Tabla 1 muestra un resumen de la calidad de los datos, los cuales fueron examinados a partir del conjunto de datos original para reducir el riesgo asociado al uso de información defectuosa. Se eliminaron un total de 97 filas y una columna que correspondían a datos erróneos y duplicados, lo que representa el 2.91% del total de los datos.

**Tabla 1.** Reporte de calidad numérico

| Ítem | Procedimiento | Detalle | Resultado |
|---|---|---|---|
| 1 | Análisis Exploratorio Descriptivo | Importación del conjunto de datos | Se ha importado 3338 filas y 28 columnas |
| | | Eliminación de columnas duplicadas | Se elimina 1 columna: "Inmobiliaria". El número de columnas resultantes es 27. |
| | | Identificación y eliminación de filas duplicadas | No existen filas duplicadas. El número resultante de filas es 3338. |
| | | Identificación de valores nulos | Se identifica el número de nulos en filas y columnas. Se mantiene aquellas que tengas al menos el 70% de datos válidos. |
| | | Eliminación de filas incompletas | Se elimina 35 filas incompletas. El número resultante de filas es 3303. |
| 2 | Normalización de los Datos | Estandarización de datos categóricos | Se elimina espacios en blanco. Se corrigen errores tipográficos. |

| | | | |
|---|---|---|---|
| | | | Se reemplazan valores. |
| | | Verificar errores en datos numéricos | Se eliminan espacios y caracteres no deseados. |
| | | Ajustar tipos de datos | Se convierte las columnas a los tipos de datos apropiados: int y float. |
| 3 | Imputación de valores faltantes | Realizar el tratamiento de datos vacíos | Se aplica estrategias como: 'mean', 'most_frecuent' y 'median', para rellenar valores faltantes en las columnas. |
| 4 | Identificación de valores atípicos | Análisis de valores atípicos<br>Eliminación de valores atípicos | Se eliminan 62 filas con valores atípicos.<br>El número resultante de filas es: 3241. |
| 5 | Verificación de la integridad y consistencia | Verificar que no existan valores vacíos | Se verifica que no queden valores nulos en filas y columnas. |
| | | Revisión de filas duplicadas | Se verifica que no existan filas duplicadas. |
| 6 | Guardado de datos limpios | Almacenar datos procesados | El nuevo conjunto de datos a exportar tiene 3241 filas y 27 columnas. |

**Preparación de los datos para PCA**

La Figura 1 ilustra el proceso de reducción de dimensionalidad. A la izquierda, se observa una matriz de datos de tamaño $n\ x\ m$, donde $n$ representa el número de observaciones (filas) y $m$ el número de variables (columnas). Tras aplicar PCA, la matriz se transforma en una nueva matriz de tamaño $n\ x\ z$, donde $z$ es el número reducido de variables resultantes (columnas) que capturan la mayor variabilidad de los datos originales. Como valor añadido, luego de aplicar PCA todas las nuevas variables son independientes una de otra.

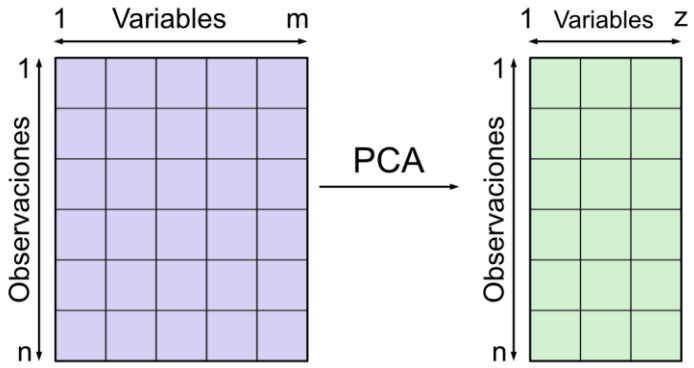


**Figura 1.** Proceso de reducción de Dimensionalidad utilizando PCA

Para poder aplicar PCA sobre un conjunto de datos específico, es importante notar que se emplea típicamente a variables numéricas. Esto se debe a que, matemáticamente, PCA se basa en la descomposición de la matriz de covarianza de las variables del proceso (Garcia-Alvarez & Fuente, 2011). Como se ha afirmado antes, la matriz de covarianza se utiliza para analizar la relación entre dos o más variables aleatorias. Por lo que, la covarianza de los elementos $n = 1$ y $m = 2$ se puede expresar como:

$$\sigma_{12} = \frac{\sum_{n=1,m=2}^{n,m}(x_1 - \bar{x})}{n} \quad (1)$$

En el conjunto de datos a utilizar, el PCA se aplica directamente a las columnas numéricas. Para las columnas categóricas de la fuente de datos, es necesario convertirlas a un formato numérico antes de poder aplicar PCA. Un método común para hacerlo es la codificación one-hot, que crea una nueva columna binaria por cada categoría única en la columna original. Este método es ampliamente utilizado para transformar variables categóricas en formato numérico (Wang et al., 2024b). Sin embargo, es importante tener en cuenta que esto aumenta la dimensionalidad de los datos al crear una nueva columna para cada categoría. Por lo tanto, después de codificar las variables categóricas, puede ser necesario aplicar nuevamente PCA para reducir la dimensionalidad.

En la Tabla 2 se muestra las 11 variables numéricas y 16 variables categóricas presentes en el conjunto de datos inmobiliarios. Estas variables se utilizaron para la aplicación del método PCA con el objetivo de reducir la dimensionalidad del conjunto inicial.

**Tabla 2**. Tipo de variables disponibles

| Ítem | Columna | Tipo de Variable |
|---|---|---|
| 1 | Código | Numérica |
| 2 | Precio | Numérica |
| 3 | Precio por metro cuadrado (terreno) | Numérica |
| 4 | Precio por metro cuadrado (construcción) | Numérica |
| 5 | N.º de pisos | Numérica |
| 6 | Área de terreno | Numérica |
| 8 | Área de construcción | Numérica |
| 8 | N.º de habitaciones | Numérica |
| 9 | N.º de baños | Numérica |
| 10 | N.º de garaje (o autos) | Numérica |
| 11 | Valoración de la propiedad del 1 al 10 | Numérica |
| 12 | Fecha de publicación | Categórica |
| 13 | Tipo de Propiedad | Categórica |
| 14 | Ciudad | Categórica |
| 15 | Ubicación | Categórica |
| 16 | Barrio | Categórica |
| 17 | Sector | Categórica |
| 18 | Descripción ubicación | Categórica |
| 19 | Estado | Categórica |
| 20 | Esquinera | Categórica |
| 21 | Tipo de calle | Categórica |
| 22 | Agua | Categórica |
| 23 | Luz | Categórica |
| 24 | Alcantarillado | Categórica |
| 25 | Cercanía a zonas comerciales | Categórica |
| 26 | Inmobiliaria | Categórica |
| 27 | Link | Categórica |

**Implementación del algoritmo para PCA**

La aplicación del método de Análisis de Componentes Principales (PCA) se llevó a cabo utilizando un algoritmo informático diseñado para reducir la dimensionalidad de los datos inmobiliarios de la ciudad de Riobamba. Este procedimiento se realizó en dos fases: primero, considerando únicamente las variables numéricas, y luego, incluyendo tanto variables categóricas como numéricas. A continuación, se describe el algoritmo aplicado y los pasos seguidos.

1. **Importar librerías necesarias:** Se importó las librerías necesarias para el análisis de datos, manipulación de matrices y visualización de resultados.
2. **Cargar los datos de entrada:** Se cargó el conjunto de datos desde un archivo Excel en un Dataframe de Pandas. El archivo de entrada tiene 3241 filas y 27 columnas.
3. **Eliminar columnas irrelevantes:** Se eliminó columnas que no aportan información relevante para el análisis. Por ejemplo, la columna código que contiene identificadores únicos para cada no es de interés por lo que queda fuera del conjunto de datos de análisis.
4. **Considerar valores numéricos:** En primera instancia, se aplicó PCA a datos numéricos, por lo que se seleccionó las variables numéricas únicamente y se eliminan las columnas de datos categóricos.
5. **Estandarizar la información:** Se estandarizó los datos para que cada característica tenga una media de 0 y varianza de 1. Es necesario estandarizar los datos de cada variable con media nula y varianza unitaria para que, en el cómputo de componentes principales, todas las variables tengan el mismo peso. Esto asegura que todas las variables sean consideradas equitativamente en el análisis (Garcia-Alvarez & Fuente, 2011).
6. **Aplicar PCA:** Se inicializó una instancia de PCA y se la aplicó a los datos estandarizados. Esto con el propósito de identificar los componentes principales y la varianza explicada por cada componente Este método es particularmente efectivo cuando los datos pueden ser representados como una nube de puntos con forma de hiperelipsoide. En tal situación, las direcciones de máxima variabilidad, que PCA busca identificar y utilizar, corresponden a los ejes principales del hiperelipsoide. Estos ejes son determinados por la matriz de varianzas y covarianzas de los datos (Rivas & Cerrillo, 2014).
7. **Obtener la varianza explicada:** Se exploró la cantidad de varianza que cada componente principal explica.

La varianza es una medida de la dispersión de los datos en un conjunto de datos. Se define como la media de la desviación de cada término de su media aritmética de datos enteros. De hecho, es casi idéntica a la desviación estándar. La covarianza siempre se mide entre 2 dimensiones y su fórmula es muy similar a la fórmula de la varianza (Mishra et al., 2017). La fórmula de la varianza y covarianza se puede expresar como:

$$Var(x) = \frac{\sum_{i=1}^{n}(X_1 - X')(X_i - X')}{(n-1)} \quad (2)$$

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(X_1 - X')(Y_i - Y')}{(n-1)} \quad (3)$$

Donde:

- $X'$ es la media aritmética de los datos $X$.
- $Y'$ es la media aritmética de los datos $Y$.
- $X - X'$ es la desviación de la observación individual con respecto a la media aritmética.
- $n$ es el número de observaciones.

8. **Obtener la varianza explicada acumulada:** Se calculó la varianza explicada acumulada por los componentes principales.
9. **Definir el umbral de la varianza explicada acumulada:** Se estableció un umbral del 95% de varianza explicada acumulada.
10. **Obtener el número óptimo de componentes principales:** Se determinó el número de componentes necesarios para explicar al menos el 95% de la varianza total.

Siendo el objetivo principal de PCA reducir la dimensionalidad de los datos, por lo general, se busca utilizar el mínimo número de componentes que sean suficientes para explicar los datos. Se han sugerido varias reglas para determinar el número óptimo de componentes principales a utilizar (Garcia-Alvarez & Fuente, 2011).

Un enfoque común es evaluar la proporción de varianza explicada acumulada y seleccionar el número mínimo de componentes donde el incremento ya no es significativo. En este nuevo sistema de coordenadas que resulta de los datos originales, la varianza mayor es conocida como componente

principal, el segundo componente es el de mayor varianza solo por debajo del primero, y así sucesivamente (Quiroga & Villalobos, 2015).

11. **Visualizar la información:** Se graficó la proporción de varianza explicada acumulada por PCA para identificar visualmente el número de componentes principales obtenido después de aplicarlo a las variables numéricas únicamente.
12. **Considerar variables categóricas y numéricas:** Del conjunto original de datos, se separó las características numéricas y categóricas. Se eliminaron las columnas categóricas que pueden generar muchas categorías únicas en la codificación, como son: "Fecha de publicación", "Ubicación", "Barrio", "Link", "Inmobiliaria".
13. **Codificar las características categóricas:** Se codificó las características categóricas utilizando OneHotEncoder. OneHotEncoder se utiliza para convertir variables categóricas en formas numéricas.
14. **Crear un nuevo dataframe:** Se combinó las características numéricas y categóricas codificadas en un solo dataframe para el análisis.
15. **Estandarizar los datos:** Se estandarizó el nuevo dataframe para que cada característica tenga media de 0 y varianza de 1.
16. **Aplicar PCA a los datos combinados:** Se aplicó PCA a las características estandarizadas para analizar la varianza explicada.
17. **Obtener la varianza explicada de los datos combinados:** Se exploró la cantidad de varianza explicada por cada componente principal.
18. **Obtener la varianza explicada acumulada de los datos combinados**: Se calculó la varianza explicada acumulada por los componentes principales.
19. **Definir el umbral de la varianza explicada acumulada de los datos combinados:** Se estableció un umbral del 95% de varianza explicada acumulada.
20. **Obtener el número óptimo de componentes principales de los datos combinados:** Se determinó el número de componentes que explican al menos el 95% de la varianza.
21. **Visualizar la información para las variables numéricas y categóricas:** Se graficó la proporción de varianza explicada acumulada para visualizar el número de componentes principales obtenidos en el umbral establecido.
22. **Evaluar el rendimiento del modelo original y PCA:** Se entrenó y evaluó un modelo de regresión lineal con los datos originales, con componentes principales de las variables

numéricas, y con los componentes principales de los datos combinados. Para ello, se definió una variable objetivo dado que la regresión lineal es un modelo supervisado. Posteriormente, se dividieron los datos en conjuntos de entrenamiento y prueba, y se estandarizó la información. La fase de entrenamiento del algoritmo, se la realizó en tres etapas.

Primero se entrenó modelo sin PCA, es decir, con el conjunto de datos original. En la segunda etapa, el modelo recibió el número óptimo de componentes principales aplicado únicamente a las variables numéricas. Por último, el modelo recibió el número óptimo de componentes principales de las variables numéricas y variables categóricas para su entrenamiento.

23. **Visualizar los resultados de evaluación:** Se graficó y comparó los resultados de MAE y $R^2$ para cada enfoque. MAE (Error Medio Absoluto) permite conocer el error promedio entre las predicciones del modelo y los valores reales, proporcionando una medida clara y fácil de interpretar de la precisión del modelo. Un valor MAE más bajo indica un modelo más preciso. $R^2$ mide la proporción de variabilidad en los datos de salida que es explicada por el modelo. Su valor varía entre 0 y 1, donde 1 indica un modelo que explica perfectamente la variabilidad de los datos y 0 indica que el modelo no explica la variabilidad en absoluto. Esto es útil porque brinda una idea de la calidad del ajuste el modelo a los datos.

## Resultados

### Aplicación de PCA en variables numéricas

Los datos sobre la varianza explicada son fundamentales para determinar cuántos componentes principales se deben utilizar en el análisis. La Figura 2 muestra la proporción de varianza explicada por cada uno de los diez componentes principales, obtenida mediante el análisis de componentes principales (PCA) aplicado al conjunto de datos inmobiliarios de la ciudad de Riobamba. Este análisis se realizó considerando únicamente las variables numéricas. Se observó que el primer componente principal explica el 27% de la varianza total, mientras que los componentes subsiguientes explican varianzas decrecientes: 17% con el segundo componente y 11% con el tercero. Este patrón indica una rápida disminución en la cantidad de varianza explicada por cada componente adicional.

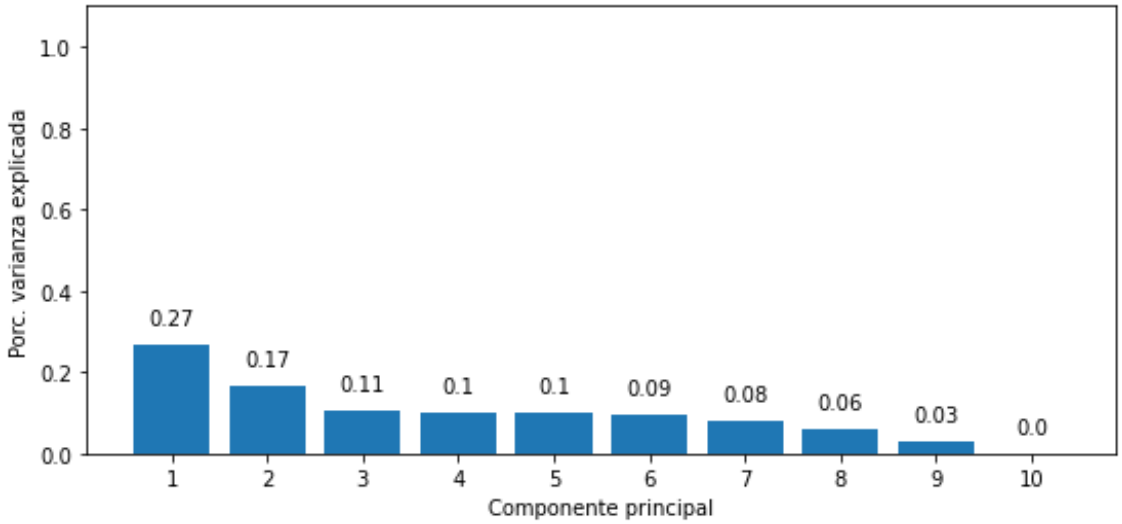


**Figura 2.** Porcentaje de varianza aplicada por cada componente

La Figura 3 presenta la proporción acumulada de varianza explicada en función del número de componentes principales. Esta gráfica permite determinar el número óptimo de componentes necesarios para explicar al menos el 95% de la varianza total. En este caso, se observó que los primeros 8 componentes principales son suficientes para alcanzar el umbral propuesto, lo que indica una reducción significativa en la dimensionalidad del conjunto de datos original sin perder una cantidad considerable de información.

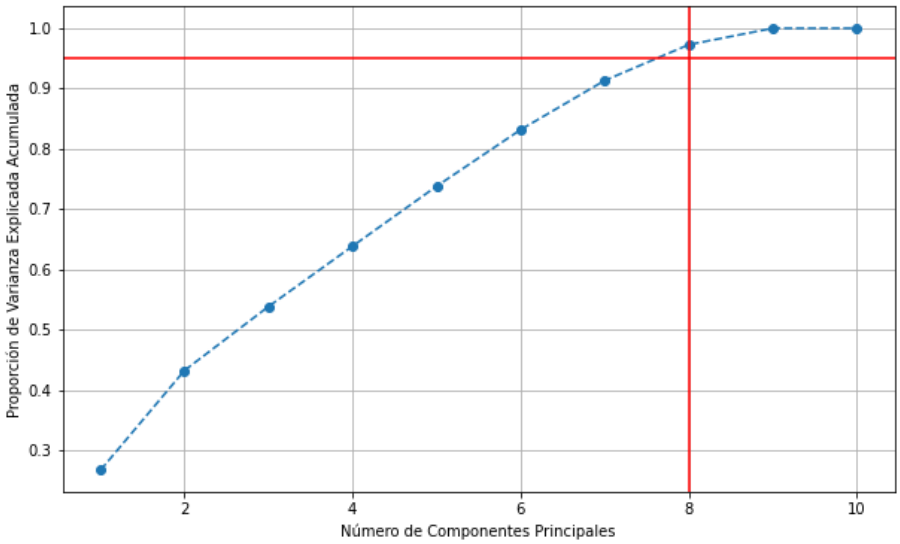


**Figura 3.** Proporción de varianza explicada acumulada

**Aplicación de PCA en variables combinadas**

Al aplicar el algoritmo PCA al conjunto de datos que incluye tanto valores numéricos como categóricos, se observaron más componentes principales en comparación con el análisis realizado únicamente sobre datos numéricos. La Figura 4 muestra la distribución de la varianza explicada por cada componente principal en este contexto. Aunque el primer componente sigue explicando la mayor parte de la varianza, la proporción de varianza se distribuye de manera más uniforme entre un mayor número de componentes. Esto sugiere una estructura de datos más compleja al incluir variables categóricas junto con las numéricas.

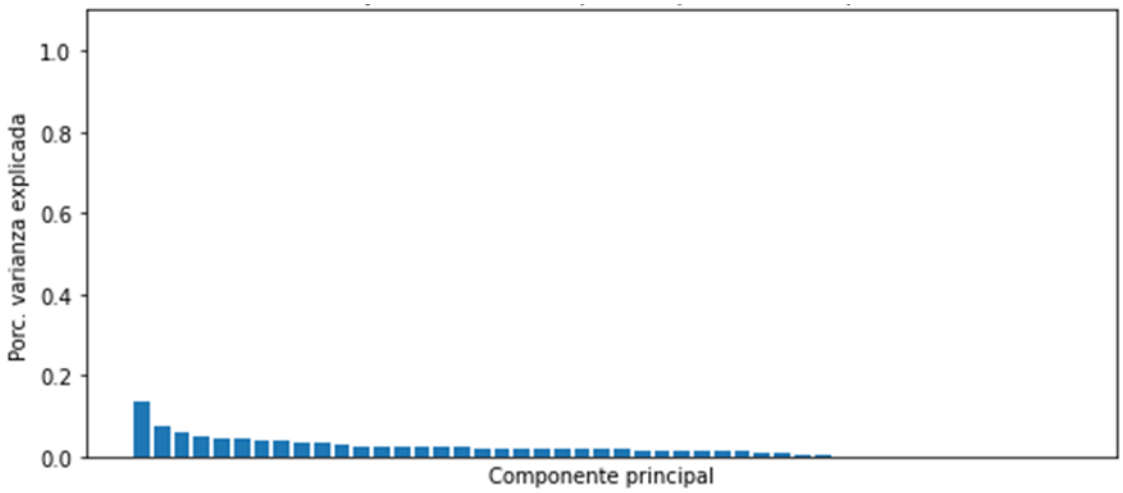


**Figura 4.** Porcentaje de varianza explicada en variables numéricas y categóricas

En la Figura 5 se puede observar la proporción de varianza explicada acumulada en función del número de componentes principales extraídos considerando las variables categóricas y numéricas. Este gráfico fue fundamental para identificar el número óptimo de componentes necesarios para capturar al menos el 95% de la varianza total en el conjunto de datos inmobiliarios de la ciudad de Riobamba. La línea roja horizontal marca el umbral del 95% de varianza explicada, mientras que la línea roja vertical indica el punto en el cual se alcanza este umbral. A partir del análisis de esta figura, se determinó que aproximadamente 30 componentes principales son suficientes para retener el 95% de la información original de los datos.

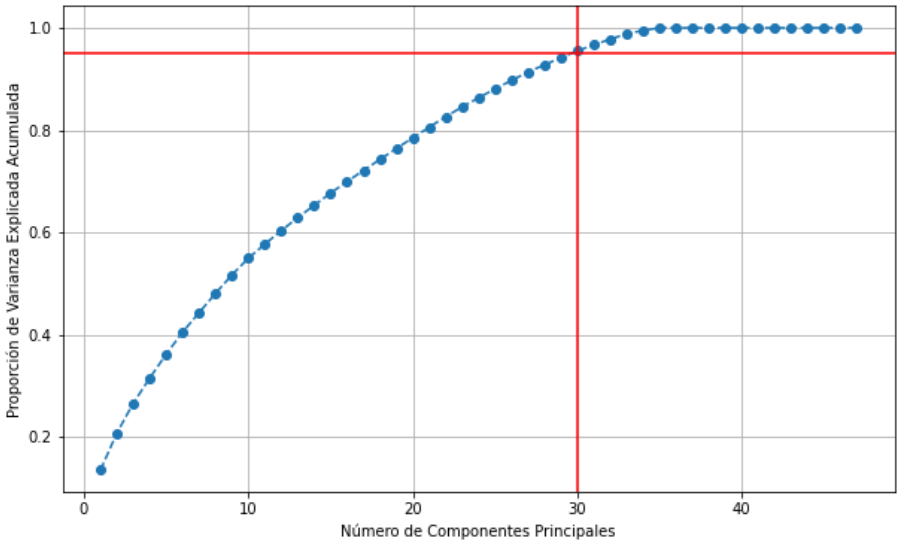


**Figura 5.** Proporción de varianza explicada acumulada en variables numéricas y categóricas

**Entrenamiento y evaluación de un modelo de regresión lineal**

En el aprendizaje supervisado, es necesario especificar una variable objetivo, ya que es la variable que el modelo intentará predecir. Para el ejemplo, 'Precio' es la variable objetivo, y se separa del resto de las características en el conjunto de datos para entrenar el modelo. La Tabla 3 muestra los resultados del rendimiento de un modelo de regresión lineal antes y después de aplicar la reducción de dimensionalidad mediante el Análisis de Componentes Principales (PCA). Se evaluaron tres casos: el conjunto de datos sin aplicar PCA, PCA aplicado solo a datos numéricos y PCA aplicado a datos numéricos y categóricos. Las métricas utilizadas para evaluar el rendimiento del modelo son el Error Absoluto Medio (MAE) y el Coeficiente de Determinación ($R^2$).

**Tabla 3.** Resultados de la valoración del rendimiento mediante regresión lineal

| Métrica | Datos Originales | PCA con datos numéricos | PCA con datos categóricos y numéricos |
|---|---|---|---|
| **MAE** | 70553.89 | 2030.04 | 54509.14 |
| **$R^2$** | -0.44 | 0.99 | -3.22 |

**Reducción del MAE**: La aplicación de PCA a los datos numéricos reduce significativamente el MAE, mejorando la precisión de las predicciones del modelo.

**Ajuste del Modelo ($R^2$)**: El modelo ajustado con PCA aplicado a datos numéricos presenta un excelente ajuste, mientras que el modelo con PCA aplicado a datos numéricos y categóricos muestra un ajuste deficiente.

**Efectividad del PCA**: PCA es altamente efectivo para mejorar el rendimiento del modelo cuando se aplica a datos numéricos. Sin embargo, la inclusión de datos categóricos en el PCA no mejora el rendimiento y puede incluso deteriorarlo.

## Conclusiones

El análisis exploratorio descriptivo, la identificación de valores atípicos, y la verificación de la integridad y consistencia de los datos son fundamentales para minimizar el riesgo de utilizar información errónea. En el caso de los datos inmobiliarios de Riobamba, la eliminación de valores atípicos y la limpieza de datos irrelevantes han permitido crear un conjunto de datos más fiable y preciso. Este proceso ha sido efectivo en mantener la integridad de la mayoría de los datos originales, con solo un 2.91% de datos eliminados, que correspondían a duplicados, datos incompletos y valores atípicos.

El nuevo conjunto de datos, limpio y de calidad, ha mostrado una mejora significativa en el rendimiento del modelo al aplicar técnicas de reducción de dimensionalidad, como PCA. La reducción del MAE (Error Absoluto Medio) de 70,553.89 a 2,030.04 al utilizar PCA con datos numéricos indica que la limpieza y preparación de los datos han tenido un impacto positivo en la calidad del modelo predictivo.

La aplicación del método PCA utilizando un algoritmo informático ha sido exitosa en la reducción de la dimensionalidad de los datos inmobiliarios de Riobamba. Al aplicar PCA solo a los datos numéricos, se logró reducir significativamente la dimensionalidad mientras se mantiene una alta proporción de la varianza explicada (95%), con un $R^2$ de 0.99 y un MAE de 2030.04. Esto demuestra que PCA es efectivo para simplificar el conjunto de datos, manteniendo la mayoría de la información relevante y mejorando la interpretabilidad y eficiencia de los modelos predictivos.

La efectividad del método PCA se ha validado a través de varios indicadores. La proporción de varianza explicada acumulada mostró que un número óptimo de componentes principales, 8 en la prueba de variables numéricas, puede explicar el 95% de la varianza total de los datos. En términos de rendimiento, el uso de PCA con datos numéricos mejoró drásticamente la precisión del modelo, obteniendo un valor MAE de 2030.04 y $R^2$ igual a 0.99. Sin embargo, al incorporar tanto datos

categóricos como numéricos, el número de componentes óptimo fue 30 para lograr un 95% de varianza total en los datos, la efectividad del PCA disminuyó, como lo indica el valor MAE elevado de 54509.14 y el $R^2$ negativo de -3.22. Esto sugiere que mientras PCA es altamente efectivo para datos numéricos, la inclusión de datos categóricos requiere una consideración cuidadosa y posiblemente técnica adicionales para mejorar el rendimiento del modelo, pues al utilizar la codificación se crea una nueva columna binaria para cada categoría única en la columna original.

## Referencias

Bodero, E., Lopez, M., Congacha, A., Cajamarca, E., & Morales, C. (2020a). Google Colaboratory como alternativa para el procesamiento de una red neuronal convolucional. Revista ESPACIOS, 41(07). https://www.revistaespacios.com/a20v41n07/20410722.html

Bodero, E., Morales, C., Congacha, A., & Ramos, C. (2022b). Técnicas de minería de datos para el análisis de la plusvalía inmobiliaria. Dominio de las Ciencias, 8(1). https://doi.org/10.23857/dc.v8i41.2531

Cheng, Z., Wang, S., Zhang, P., Wang, S., Liu, X., & Zhu, E. (2021). Improved autoencoder for unsupervised anomaly detection. International Journal of Intelligent Systems, 36(12), 7103-7125. https://doi.org/10.1002/int.22582

Cruz, E., González, M., & Rangel, J. (2022). Técnicas de machine learning aplicadas a la evaluación del rendimiento y a la predicción de la deserción de estudiantes universitarios, una revisión. Prisma Tecnológico, 13(1), Article 1. https://doi.org/10.33412/pri.v13.1.3039

Dray, S. (2008). On the number of principal components: A test of dimensionality based on measurements of similarity between matrices. Computational Statistics & Data Analysis, 52(4), 2228-2237. https://doi.org/10.1016/j.csda.2007.07.015

Garcia-Alvarez, D., & Fuente, M. (2011). Estudio comparativo de técnicas de detección de fallos basadas en el Análisis de Componentes Principales (PCA). Revista Iberoamericana de Automática e Informática Industrial RIAI, 8(3), 182-195. https://doi.org/10.1016/j.riai.2011.06.006

Gonzalez, V., Conde, G., & Muñoz, A. (2021). Análisis de Componentes Principales en presencia de datos faltantes: El principio de datos disponibles: Principal Components Analysis in the presence of missing data: the principle of available data. Scientia et Technica, 26(2), Article 2. https://doi.org/10.22517/23447214.20591

Jiménez, J., Herrera, A., & Rojas, W. (2002). Fuentes de Varianza e Indices de Varianza Explicada en las Ciencias del Movimiento Humano. Pensar en Movimiento: Revista de Ciencias del Ejercicio y la Salud, 2(2), Article 2. https://doi.org/10.15517/pensarmov.v2i2.398

Kurita, T. (2019). Principal Component Analysis (PCA). In Computer Vision: A Reference Guide (pp. 1-4). Springer International Publishing. https://doi.org/10.1007/978-3-030-03243-2_649-1

Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., Panda, S., & Laishram, M. (2017). Multivariate statistical data analysis-principal component analysis (PCA). International Journal of Livestock Research, 7(5), 60-78. https://acortar.link/NVFGPE

Mostofi, F., Toğan, V., & Başağa, H. (2022). Real-estate price prediction with deep neural network and principal component analysis. Organization, Technology and Management in Construction, 14(1), 2741-2759. https://doi.org/10.2478/otmcj-2022-0016

Pernice, S. (2024). El problema de la reducción dimensional. Análisis de Componentes Principales (PCA). Revista Mutis, 14(1), Article 1. https://doi.org/10.21789/22561498.2057

Quiroga, C., & Villalobos, A. (2015). Análisis del comportamiento bursátil de las principales bolsas financieras en el mundo usando el análisis multivariado (análisis de componentes principales PCA) para el periodo de 2011 a 2014 (Analysis of Stock Market Behavior of the Major Financial Exchanges Worldwide Using Multivariate Analysis (Principal Component Analysis PCA) for the Period 2011 to 2014). Revista CEA, 1(2). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3520246

Rivas, J., & Cerrillo, S. (2014). Un Algoritmo Genético para Selección de Kernel en Análisis de Componentes Principales con Kernels. Investigación Operacional, 35(2), Article 2. https://revistas.uh.cu/invoperacional/article/view/4713

Sarker, I. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Wang, Z., Zhang, G., Xing, X., Xu, X., & Sun, T. (2024a). Comparison of dimensionality reduction techniques for multi-variable spatiotemporal flow fields. Ocean Engineering, 291, 116421. https://doi.org/10.1016/j.oceaneng.2023.116421

Wang, Z., Li, W., & Tang, Z. (2024b). Enhancing the genomic prediction accuracy of swine agricultural economic traits using an expanded one-hot encoding in CNN models1. Journal of Integrative Agriculture. https://doi.org/10.1016/j.jia.2024.03.071