Universidad Nacional Agraria La Molina

Facultad de Economía y planificación

Estadística Informática



**Docente:**

Meza Rodriguez Aldo

**Integrantes:**

Juárez Castro, Andre Saul – 20200396

Medrano Alania, Sebastian Rodrigo – 20211821

Quispe Cuadros, Arthur Jesus Abrahan – 20211827

Vargas Cespedes, Jose Enmanuel – 20191317

**Presentación:**

En el presente proyecto se desarrolla un modelo predictivo para estimar el precio de venta de viviendas en Lima Metropolitana, utilizando técnicas de aprendizaje automático, específicamente redes neuronales profundas. El objetivo es entender cómo las características estructurales y geográficas de los inmuebles influyen en su valor de mercado, en un contexto urbano marcado por una alta desigualdad económica y una fuerte heterogeneidad en los precios de las propiedades.

**1. Selección y exploración de datos**

**Se seleccionaron las variables:**

**Variables independientes:**

Dormitorios Baños Garajes M² edificados Tipo de Propiedad Estado del inmueble Zona\_macro

Variable dependiente: **Precio**

#### **Dataset**

El dataset utilizado está compuesto por **434 registros** de propiedades en venta ubicadas en Lima Metropolitana, Perú. Cada registro representa un inmueble (casa o departamento) y contiene variables estructurales y geográficas que podrían influir en su valor de venta. Las variables incluidas son:

* **Dormitorios** (número de habitaciones)
* **Baños** (cantidad de baños)
* **Garajes** (número de cocheras)
* **M² edificados** (área construida)
* **Tipo de Propiedad** (Casa o Departamento)
* **Estado del inmueble** (Buen estado, Excelente estado, etc.)
* **Zona\_macro** (zona geográfica de Lima: Norte, Sur, Centro, etc.)
* **Precio** (valor de venta en soles peruanos)

Estas variables permiten explorar las características que más influyen en el precio de venta de los inmuebles y construir modelos predictivos sólidos.

#### **1.2 Fuente de los Datos**

Los datos fueron recolectados mediante técnicas de **web scraping**, específicamente desde los siguientes portales inmobiliarios:

* 🏠 [Infocasas - Casas en Venta en Lima](https://www.infocasas.com.pe/venta/casas/lima)
* 🏢 [Infocasas - Departamentos en Venta en Lima](https://www.infocasas.com.pe/venta/departamentos/lima)

Ambos portales ofrecen listados públicos de inmuebles en venta, con detalles clave que fueron extraídos automáticamente usando scripts de scraping desarrollados en Python. Posteriormente, la información fue depurada, unificada y transformada para su análisis.

**2. Preprocesamiento y análisis exploratorio**

Caso de la variable zona

La variable Zona presenta una alta cantidad de categorías únicas (≈ 40), muchas de las cuales tienen muy pocos registros. Para evitar una explosión de dimensionalidad al aplicar codificación One-Hot se creó una nueva variable Zona\_macro agrupando los distritos de Lima Metropolitana en macrozonas geográficas: Lima Norte, Sur, Este, Centro, Callao y Otros.

Esto permite reducir la cardinalidad de la variable Zona, evitar la creación excesiva de variables dummy, y mantener una representación territorial coherente con la estructura urbana y socioeconómica de la ciudad.

La variable Estado del inmueble describe la condición en la que se encuentra la propiedad (por ejemplo, “Buen estado”, “Excelente estado”, “Regular”, etc.). Durante la exploración de esta variable se identificó que la mayoría de registros se concentraban en dos categorías principales: **“Buen estado”** y **“Excelente estado”**, mientras que el resto de valores eran poco frecuentes o poco informativos.

Para evitar la creación de muchas columnas dummy poco representativas (producto del one-hot encoding), y reducir la **sparsidad** y el **ruido**, se decidió **agrupar todas las categorías diferentes a “Buen estado” y “Excelente estado” bajo una nueva categoría llamada “Otro”**.

#### **2.1 Limpieza y Preparación de los Datos**

Una vez recolectados los datos mediante scraping, se procedió a depurar el dataset. Se eliminaron columnas irrelevantes como identificadores únicos y se identificaron registros con valores atípicos extremos en la variable Precio. Debido a la alta dispersión en los precios de viviendas en Lima, se aplicó una **winsorización** al 1% inferior y superior para limitar la influencia de valores anómalos sin eliminar información valiosa. Además, se aplicó una transformación matemática (raíz cuadrada) al precio para estabilizar la varianza y mejorar el desempeño de los modelos. Los datos numéricos fueron escalados con RobustScaler para reducir la sensibilidad a outliers, mientras que las variables categóricas se codificaron mediante one-hot encoding.

**2.2 Visualización y Análisis de Características**

Se realizaron visualizaciones exploratorias que permitieron entender la distribución de las variables y su relación con el precio. Se identificó, por ejemplo, que propiedades ubicadas en Lima Centro o en estado “Excelente” tendían a tener precios superiores al promedio. Asimismo, se observaron relaciones positivas entre el número de dormitorios, metros cuadrados construidos y el valor de venta. Estas visualizaciones sirvieron como punto de partida para validar la importancia potencial de cada variable en la predicción del precio.

**3. Implementación de modelos de redes neuronales artificiales**

Con los datos preprocesados, se implementaron dos modelos de redes neuronales artificiales para predecir el precio de venta de viviendas. Ambos modelos utilizaron como entrada las mismas variables estructurales y categóricas transformadas, pero se desarrollaron utilizando dos bibliotecas distintas: Scikit-learn y PyTorch. Esto permitió comparar el desempeño de una red neuronal de alto nivel (Scikit-learn) con una red personalizada y optimizable en bajo nivel (PyTorch).

#### **3.1 Modelo 1: Red neuronal con Scikit – Learn**

Se implementó un MLPRegressor utilizando la biblioteca Scikit-learn. El modelo fue entrenado con un pipeline completo de preprocesamiento y se ajustaron los hiperparámetros mediante RandomizedSearchCV, considerando distintas estructuras de red, tasas de aprendizaje y niveles de regularización. Aunque el modelo logró capturar parte de la variabilidad del precio, su desempeño se vio limitado por la alta dispersión de la variable objetivo. El mejor modelo alcanzó un coeficiente R² de aproximadamente **0.52**, con un error medio en soles razonable dado el rango de precios observados.

#### **3.2 Modelo 2: Red neuronal con PYtorch**

A diferencia del modelo anterior, en PyTorch se definió manualmente la arquitectura de la red, utilizando tres capas ocultas con funciones de activación ReLU. Se implementó un ciclo de entrenamiento personalizado con descenso del gradiente, optimización con Adam y función de pérdida MSE. Además, se desarrolló un sistema de búsqueda de hiperparámetros (número de neuronas, función de activación, regularización) para seleccionar la mejor arquitectura. Este modelo logró un desempeño comparable al de Scikit-learn, y permitió un mayor control sobre el proceso de entrenamiento.

#### **3.3 Limitaciones**

Ambos modelos enfrentaron limitaciones debido a la **alta varianza y asimetría** en la distribución del precio de los inmuebles. Aunque se aplicaron técnicas de transformación y escalado, la heterogeneidad socioeconómica de Lima genera un patrón de precios difícil de modelar. Adicionalmente, la ausencia de variables clave como ubicación exacta (distrito, calle), antigüedad del inmueble, o características adicionales (ascensor, piscina, vigilancia, etc.) restringe el poder predictivo del modelo. A pesar de ello, se lograron resultados aceptables que permiten obtener una estimación aproximada del precio.

**4. Conclusión**

Este proyecto demuestra cómo aplicar redes neuronales artificiales para abordar un problema real: la estimación del precio de viviendas en una ciudad con características socioeconómicas complejas como Lima. Mediante un proceso riguroso de recolección, limpieza, transformación y modelado, se implementaron dos enfoques basados en Scikit-learn y PyTorch. Si bien los modelos presentan limitaciones derivadas de la variabilidad de los datos, se logró alcanzar un R² superior al 0.5, lo cual representa un resultado positivo considerando el nivel de detalle disponible. En trabajos futuros, se recomienda enriquecer el dataset con variables geográficas y cualitativas adicionales para mejorar la precisión de las predicciones.